Against Sainsbury and Tye's Originalism

A Critical Investigation of an Originalist Theory of Concepts and Thoughts

Sara Kasin Vikesdal

Thesis presented for the degree of

MASTER OF PHILOSOPHY

Supervised by Professor Carsten Hansen

Department of Philosophy, Classics, History of Art and Ideas

Faculty of Humanities

University of Oslo

Spring 2015

Abstract

In *Seven Puzzles of Thought and How to Solve Them* Sainsbury and Tye defend a version of originalism, the view that concepts are to be individuated by way of their origins. A consequence of their account is that concepts that are semantically distinct may nonetheless be of the same type (and vice versa). In this thesis I argue that a result of this commitment is that their account fails as a general theory of concepts and thoughts. I show by appeal to a thought experiment that Sainsbury and Tye's originalism cannot provide a general account of the cognitive role of concepts and thoughts: the theory fails to explain certain cases of rationality. Further, I show that originalism fails at the specific task of solving three classical puzzles within the philosophy of mind and language; puzzles the solution of which is the *raison d'être* of originalism.

Against Sainsbury and Tye's Originalism: A Critical Investigation of an Originalist Theory of Concepts and Thoughts

http://www.duo.uio.no/

Acknowledgments

Foremost, I wish to express my sincere gratitude to my supervisor, Carsten Hansen, for all help and support I've received while working on my thesis. Especially I wish to thank him for valuable feedback and for his selfless use of time. I also would like to thank the Centre for the Study of Mind in Nature (CSMN) for granting me a stipend; the RAR (Representationalism or Anti-Representationalism?) project for letting me attend the workshop on acquaintance and the conference on perspectives on rationality; and Kim Phillips Pedersen for indispensible discussions, for proofreading and for love and support.

Contents

Introduction

Thoughts are central to our reasoning and action. For instance, yesterday I went to the store in order to buy milk. I did this because I had a desire for milk and because I believed that going to the store would result in me getting milk. When desiring milk I formed the thought *I want milk.* I also formed the thought *going to the store will result in me getting milk.* From this I concluded that going to the store would get me what I wanted. I was rational in doing so, because it follows from my desire and belief that, if they are true, going to the store will result in me getting what I want. This piece of practical reasoning then resulted in me going to the store. A central feature of the two thoughts seems to be that they are directed towards the same thing, namely milk. Had I desired milk and thought that going to the store would result in me getting juice, I would not be rational in going to the store on the basis of these thoughts alone. Further, if I had the desire that produced the thought *I want milk* and the belief *going to the store will result in me getting white liquid produced by the mammary glands of cows* I would only conclude that going to the store would get me what I want if I also had the belief that milk is white liquid produced by the mammary glands of cows. If I did not have this belief, I would not be rational in concluding from my desire and belief alone that going to the store would get me what I want. Even though the two thoughts are directed towards the same object, milk, they seem to play different roles in my cognition: I may think I want milk and at the same time think that I do not want white liquid produced by the mammary glands of cows, if I lack the belief that the two thoughts are about the same object. Then it seems that what is important for the cognitive significance of thoughts, and hence my being rational, is not the object itself but rather how I relate to the object in question. What does this tell us about the nature of thoughts? It seems that a general theory of thoughts must be able to explain why I acted the way I did when going out to buy milk, and also why I was rational in doing so. What view about the metaphysics of thoughts do we need for this explanatory task? What is a

thought, and what makes a thought the type of thought that it is instead of some other thought? In virtue of what does a thought have its specific content?

In this thesis I will discuss one of the most recent theories suggesting an answer to these questions; *originalism*. The theory attempts to answer the questions by giving an account of concepts, which, according to a widely held view, are the constituents of thoughts. Given that thoughts are composed of concepts, one must understand the nature of concepts in order to understand the nature of thoughts. In *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts* (2012) (henceforth *Seven Puzzles of Thought*), Mark Sainsbury and Michael Tye take concepts deployed in thought to be analogous to words in a language. They point to what they take to be central properties of words: Words are vehicles of meaning - they express, but are not themselves linguistic meaning. Further, two words that express the same meaning (i.e. true synonyms) may nonetheless be of distinct types, if they have different etymological history. Also, two words that differ in meaning may be of the same type if they share etymological history. Hence, according to Sainsbury and Tye, words are to be individuated historically. Now, originalists take concepts to share these central properties of words: concepts are *vehicles of content*; concepts express contents but they are not themselves content. Further, two concept tokens may be of distinct types even if they express the same content, if they have different historical origins. Also, two concept tokens may be of the same type even if the express distinct contents, if they have the same historical origin.[1] Hence, Sainsbury and Tye take concepts to be individuated by way of their origins: Two concepts are of the same type if and only if they have the same origin. This is the key originalist claim. Now since concepts are the constituents of thoughts, the individuation of thoughts also depends on the origins of concepts. Understanding thoughts this way, Sainsbury and Tye claim, will help us answer the central questions above.

The main motivation behind originalism is that it allegedly solves seven of the main puzzles within the philosophy of mind. These are puzzles concerning the cognitive significance of thoughts – what role a given thought plays in cognition - which any theory of thoughts and concepts must account for. If a theory can in fact give a satisfactory solution to all of these puzzles at once, we have reason to think that the theory is true. I will, however, argue that Sainsbury and Tye fail to give a satisfactory explanation of three of the seven puzzles they set

[1] I will give several examples of these possibilities throughout the thesis.

out to solve. If I am right that their formulation of originalism fails to solve certain of the puzzles, it seems that their account cannot provide a general theory of thoughts and concepts.

As mentioned, a general theory of thoughts must provide an explanation of cognitive significance. What role thoughts play in mind is essential for understanding rationality. According to originalism, cognitive significance depends on the vehicles of content. Since the thoughts *I want milk* and *I want white liquid produced by the mammary glands of cows* contain distinct concepts, they play different roles in cognition, even though the thoughts refer to the same object. This explains why I may be rational in holding one of them to be true, while rejecting the other. As we shall see, however, a consequence of originalism as stated in *Seven Puzzles of Thought* is that an individual may come to use the same concept to express distinct and contradictory contents. I will put forth a thought experiment that shows that if one takes cognitive significance to depend on vehicles of content and at the same time allows for concept tokens of the same type to have contradictory content, this threatens one's ability to explain rationality. Since Sainsbury and Tye's (2012) main concern is with the seven puzzles, they give no explicit indication of what a general theory of cognitive significance would look like on their account – they only offer solutions to the specific problem cases. I therefore offer suggested routes on their behalf. The thought experiment I put forth comes in two versions, in order to show that the two routes available to Sainsbury and Tye both fail to explain certain cases of rationality. If Sainsbury and Tye's account of cognitive significance fails to explain rationality, we have further reasons to reject their originalist theory as a general theory of concepts and thoughts.

The Structure of the Thesis

The aim of this thesis is to show that Sainsbury and Tye's formulation of originalism fails as a general theory of concepts and thoughts.[2] Sainsbury and Tye hold that the key claims of originalism are compatible with various views about different aspects of mental content.[3] They do, however, present what they take to be the correct view of such matters. It is

[2] Since originalism is a relatively recent theory (2011 & 2012), it has not been much discussed. The current literature on originalism includes Millikan (2011), Recanati (2012, 244-45), Horwich (2014), and Hedger (forthcoming).

[3] For instance, they take their claims to be compatible with both externalism and internalism about mental content. I shall return to this.

Sainsbury and Tye's specific formulation of originalism with these commitments included that I will consider in this thesis. The structure of the thesis is this: I start by laying out some preliminaries to originalism. Despite being a new theory of how to individuate concepts and thoughts, originalism, as defended by Sainsbury and Tye, adopts many features of classical theories in philosophy of mind and language. A brief introduction to earlier theories will thus be helpful in understanding originalism. In chapter 2, I give a detailed overview of originalism as stated in *Seven Puzzles of Thought*. I will point to those features of originalism that are familiar from classical theories within philosophy of mind and philosophy of language and also those aspects of originalism that set the view apart from earlier theories. I will offer minor critical comments along the way, but the main discussion I leave for chapter 4 and 5. In chapter 3 I present the seven puzzles of thought and the originalist solution to these puzzles given in *Seven Puzzles of Thought*. In chapter 4, I will argue that Sainsbury and Tye fail to solve three of the puzzles their theory is advanced to solve. The reason given for believing originalism to be true is that it solves the seven puzzles; if it fails to do so, we have little reason to think that originalism can provide a general theory of concepts and thoughts. In chapter 5 I challenge Sainsbury and Tye's account further. Here, I set out a thought experiment that shows that originalism fails to account for cognitive significance and - as a consequence of this - rationality. A central explanatory role of thoughts is to account for rationality; a theory of thoughts that fails to account for this explanatory role of cannot provide a full understanding of the nature of such entities. In chapter 6 I consider possible solutions to my thought experiment on behalf of the originalists. I argue that these solutions are unsuccessful and that originalism, as stated in *Seven Puzzles of Thought*, thus fails to provide an answer to my criticism. In chapter 7, I briefly present an alternative to originalism; François Recanati's theory of mental files (2012). Recanati agrees with Sainsbury and Tye that the role of thoughts in cognition depends on the vehicles of content. He disagrees with originalism that the vehicles of content are to be individuated in terms of their origin; Recanati takes such vehicles to be individuated by their function. I will show that the mental file framework of Recanati is better suited to solve the seven puzzles of thought than Sainsbury and Tye's originalism (and in particular the problems I offered in Chapter 4). This framework also avoids the problems posed for Sainsbury and Tye by my thought experiment. Hence it seems that if one wants to hold that cognitive significance relies on the vehicles of content, Recanati's theory is favourable to Sainsbury and Tye's originalism, since Recanati's account avoids the problems posed for originalism in this thesis. The purpose of comparing originalism to Recanati's theory is not to give a general argument in favour of the latter – the

focus of this thesis is on originalism – but rather to show what is wrong with the originalist framework advocated by Sainsbury and Tye, and also to show that some of the compelling features of originalism can be maintained when cast within another framework. There might of course be other features of Recanati's theory worthy of criticism, but validating Recanati's theory as a whole is beyond the scope of this thesis; I will only discuss to what extent Recanati's theory can solve the specific problems posed for Sainsbury and Tye's originalism. Finally, in chapter 8, I conclude by saying that Sainsbury and Tye's theory fails to provide a general account of cognitive significance and rationality. The reason is that the theory fails to explain the puzzles it sets out to solve, and that it encounters further problems – problems that do not rise for Recanati's theory of mental files – when explaining certain cases of rationality. The upshot is that, at least on the basis of the cases discussed in this thesis, originalism should be abandoned in favour of competing theories of concepts and thoughts.

Chapter 1

Preliminaries

A common view is that thoughts are like sentences in the language of thought. On this view, thoughts consist of concepts governed by syntax. Hence, in order to understand the nature of thoughts one must understand the nature of concepts. Consider the thought *penguins are birds;* this thought consists of the concepts *penguin* and *bird* structured in a certain way. In addition to having a syntactic structure, thoughts also have semantic features; they express a meaning or content. Thoughts have truth conditions; for instance, the thought *penguins are birds* is true just in case penguins are birds. In this way, thoughts can be said to represent the world to be in one way or another. Central questions within philosophy of mind are these: What is the nature of the content of thoughts and concepts? How do the syntactic features of thoughts and concepts relate to the semantic features of such entities, and how are we to understand the relation between thoughts and their referents? In this chapter I will give a brief overview of some of the main camps in the debate about the individuation of concepts and thoughts: *direct reference theories* and *descriptivism*.[4] Such theories give rise to many features of originalism and hence, getting an overview of earlier theories will be helpful in understanding originalism.

[4] Many of the puzzles occupying philosophers trying to understand the nature of thoughts, were originally introduced as puzzles about language. It has, however, become common to transpose classical problems in philosophy of language to philosophy of mind. Since this thesis concerns thoughts rather than language I offer simple modifications of the classical views in order to apply them to thoughts.

1.1 Classical Theories of Concepts and Thoughts

A traditional approach to the nature of concepts is to hold that such entities must be understood in terms of their contents. There are two central questions regarding the content of concepts: (1) What is the content of concepts, and (2) in virtue of what does a concept have its reference (if any). In answering the first question, a common view is that the content of concepts simply is their reference. For instance, the content of the concept *Aristotle* is Aristotle himself. This view dates back to John Stuart Mill's (1843) theory of meaning of proper names. Theories adopting a Millian framework for concepts hold that the syntactic features of concepts are just names for objects or groups of objects, and the relation between the syntax and referent is direct, and not mediated by an object's properties. Such theories are often labelled *direct reference theories*.

One problem for direct reference theories, such as Millianism, is that some concepts such as *Pegasus* and *Vulcan* lack referents. Even so, the concepts seem to contribute to the content of thoughts. The thought *Pegasus is a winged horse* seems to play a different role in cognition than the thought *Vulcan is a winged horse*: It is one thing to believe that Pegasus is a winged horse, another to believe that Vulcan is a winged horse. How can this difference in cognitive significance be accommodated within a direct reference framework? It seems that one cannot explain empty concepts contributing to the content of thoughts on a Millian framework, according to which the content just is the objects referred to, given that there are no such objects in these cases.

Millianism is further challenged by Frege's observation that co-referential concepts can play different roles in cognition (Frege 1892). Consider the following: The Ancient Babylonians used the concept *Hesperus* to pick out the brightest star visible in the night sky and *Phosphorus* to pick out the brightest star visible in the morning. Unbeknownst to the Ancient Babylonians, *Hesperus* and *Phosphorus* actually refer to the same heavenly body, namely the planet Venus. According to classic Millianism, then, the concepts *Hesperus* and *Phosphorus* should have the same content, since they refer to the same object. However, the Ancient Babylonians had the belief that Hesperus is Hesperus, but they did not believe that Hesperus is Phosphorus. In the evening they believed that Hesperus was visible, but did not think that Phosphorus was visible. Hence, it seems like *Hesperus* and *Phosphorus* play distinct roles in cognition. In both cases the thought believed has the same content as the thought denied, on a

classical Millian framework, and classical Millians have no further resources to account for the difference in cognitive significance.

Further, the thought *Hesperus is Hesperus* is trivial, in that is does not add anything to our knowledge. The thought *Hesperus is Phosphorus*, on the other hand, is informative, in that it potentially adds something to our knowledge about the world. Since the two thoughts differ in the level of informativeness they must differ in cognitive significance. When the Ancient Babylonians discovered that Hesperus is Phosphorus they did not simply make a cognitive discovery – they did not just discover that their terms had the same meaning – according to Frege (1892, 56), they also made an empirical discovery. Furthermore, the knowledge that Hesperus is Phosphorus allows one to make inferences about the world that one would not be rational in making if one only knows that Hesperus is Hesperus. For instance, after making the discovery, the Ancient Babylonians concluded that the same heavenly body is visible twice a day. They were justified in making this inference since they knew that (i) Hesperus is visible in the evening, (ii) Phosphorus is visible in the morning, and (iii) Hesperus is Phosphorus. Without the knowledge of (iii), the conclusion that the same heavenly body is visible twice a day would not be justified, since the conclusion does not follow from (i) and (ii) alone. The question raised for direct reference theories, then, is how we can explain the difference in level of informativeness in the thought *Hesperus is Hesperus* and *Hesperus is Phosphorus,* when *Hesperus* and *Phosphorus* have the same content on such views. A full account of the nature of thoughts must account for this, since, as we have seen, it is part of the explanatory role of thoughts to account for rationality.

An alternative to direct reference theories that emerged from Frege's observations is *descriptivism.* This is the view that we can only be related in thought to objects through their instantiated properties. According to descriptivists, our knowledge of objects is mediated by our knowledge of their properties.[5] On this view, the relation between the syntactic features of concepts and their referents is not direct, in the way Millians hold. Instead, the content of concepts is determined by way of a set of associated descriptions. Take for instance the concept *Aristotle*: typically the associated descriptions are something like 'the teacher of Alexander the Great', 'Greek philosopher' etc. According to descriptivism, such descriptions determine the referent of concepts. The referent of *Aristotle* is whatever x satisfies all the associated descriptions, namely Aristotle himself. Hence, descriptivism provides an answer to

[5] Other proponents of various versions of the descriptivist theory include Russell (e.g. 1911), Searle (e.g. 1958, 1983), and Strawson (e.g. 1959).

both question (1) about what the content of concepts is, and also question (2) about in virtue of what a concept has its reference: as Kallestrup explains, according to descriptivism "meaning is fully determined by competent speakers' [or thinker's] mental associations. On this view, meaning is firmly in the mind of competent speakers [or thinkers]. But descriptivism is also a theory of reference in that descriptive content is what determines reference: a particular object is the referent of a referring term if and only if that object satisfies all the associated description" (Kallestrup 2012, 14).[6]

Although this is controversial, Frege is often taken to be one of the first descriptivists[7]. Frege introduces the notion of *sense* as part of an expression's semantic value, in addition to reference. While classical Millianism has a single levelled semantics, Frege advocates a two-levelled semantics consisting of reference and sense. A sense expresses a *mode of presentation.*[8] On one reading of Frege, senses are sets of associated descriptions.[9] Hence, the semantic value of modes of presentation depends on how individuals conceive the object referred to.[10] This framework explains why co-referential terms, such as *Hesperus* and *Phosphorus,* can play different roles in cognition:

> If the sign 'a' is distinguished from the sign 'b' only as object (here, by means of its shape), not as sign (i.e. not by the manner in which it designates something), the cognitive value of a=a becomes essentially equal to that of a=b, provided a=b is true. A difference can arise only if the difference between the signs corresponds to a difference in the mode of presentation of that which is designated (Frege 1892, 57).

In the case of *Hesperus* and *Phosphorus,* the expressions refer to the same object but they express different ways of conceiving the planet Venus. The sense of *Hesperus* is something like 'the evening star', while the sense of *Phosphorus* is something like 'the morning star'. The difference in sense explains the two concepts playing different roles in cognition: the propositions expressed by thoughts employing them differ, contrary to the Millian view. This

---

[6] Since, on this view, the content of concepts is a thinker's associated descriptions, the content of concepts and thoughts depend on intrinsic features of the thinker alone. This is often labelled *internalism.* I return to this view in 3.2.

[7] Cf. Burge 2005.

[8] Even though Frege says that senses *express* modes of presentations, I will follow the tradition of using *sense* and *mode of presentation* interchangeably when laying out the descriptivist position.

[9] It is possible that Frege would also allow non-descriptive senses (Burge 1979a). But for simplicity (it is not relevant for the debate about originalism), I will follow Kripke (1980) and assume that Frege was a descriptivist.

[10] Frege distinguishes sense from *idea.* Sense is objective in a way that ideas are not: "The reference of a proper name is the object itself which we designate by its means; the idea, which we have in that case, is wholly subjective; in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself" (Frege 1892, 59). While ideas are individual, and often coloured by personal idiosyncrasies, senses can be shared amongst different individuals.

also explains the difference in the level of informativeness in the two identity statements *Hesperus is Hesperus* and *Hesperus is Phosphorus*: the knowledge that the morning star is the morning star is trivial whereas learning that the morning star is the evening star is an important empirical discovery. Also, when knowing that the morning star is the same as the evening star we can make inferences about the world that we would not be justified in making when only knowing that the morning star is the same as the morning star. This explains why the Ancient Babylonians were only justified in concluding that the same heavenly body is visible twice a day after making the discovery that Hesperus is Phosphorus: the thought *Hesperus is Hesperus* has a different content than *Hesperus is Phosphorus* even though the thoughts contain co-referential concepts, since *Hesperus* and *Phosphorus* have different senses.

The descriptivist framework can also explain how empty concepts can play an interesting cognitive role. Even thought lacking a referent, the concept *Pegasus* has a sense/mode of presentation (i.e. winged horse). The mode of presentation accounts for the cognitive significance of empty expressions. To Frege a thought just is sense (1892, 62). Hence, in the case of empty thoughts, a thought is the same as it would be if it actually had a referent. Further, since *Pegasus* and *Vulcan* have distinct modes of presentation, they play different roles in cognition. For instance, on the one hand, when a person has the belief that Pegasus is a planet she stands in a relation to a proposition containing a mode of presentation of Pegasus, i.e. a set of descriptions she associates with the concept, where the property of being a planet is predicated on whatever satisfies the descriptions. On the other, when the person believes that Vulcan is a planet she stands in a relation to a proposition containing her mode of presentation of Vulcan, i.e. her associated descriptions, where the property of being a planet is predicated on whatever satisfies the descriptions. Since she associates different descriptions with the two concepts, the beliefs express different propositions. This explains why the beliefs have different truth conditions and how *Pegasus* and *Vulcan* can play distinct cognitive roles despite both being empty terms.

Even though descriptivism seemingly provides compelling solutions to the problems encountered by direct reference theories, the view faces problems of its own. According to descriptivism, if the only thing associated with the concept *Aristotle* is the property of being the teacher of Alexander the Great, when using the concept one refers to every x that satisfies that description. However, as Kripke (1980) points out, Aristotle might not have been the

teacher of Alexander the Great. That is, there are possible worlds in which someone else than Aristotle had the role of being Alexander's teacher. Hence, the description 'the teacher of Alexander the Great' might have failed to pick out Aristotle. Kripke introduces the notion of *rigid designators*. A rigid designator is a term that picks out the same object in all possible world in which that object exists, and fails to pick out anything in those worlds where the object does not exist: "Let's call something a *rigid designator* if in every possible world it designates the same object, *a nonrigid* or *accidental designator* if that is not the case" (Kripke 1980, 48). Transposed to concepts, the reference of concepts is modally stable; the concept *Aristotle* picks out Aristotle in every possible world, including those in which Aristotle has a different name. Consequently, the thought *Aristotle might not have been Aristotle* is false. However, 'the teacher of Alexander the Great' is nonrigid, since the description applies to someone else than Aristotle in possible worlds in which Aristotle was not the teacher of Alexander the Great. Hence, the thought *Aristotle might not have been the teacher of Alexander the Great* is true. But if the sense of *Aristotle* just is 'the teacher of Alexander the Great' the two thoughts have the same sense. Hence, the thoughts should play the same role in cognition according to descriptivists, since they hold that cognitive significance is determined by the set of associated descriptions. This cannot be the case, however, since one is true while the other is false. Hence, descriptivists face problems when accounting for the cognitive difference of thoughts expressing the same mode of presentation.[11]

If one takes this criticism of descriptivism to be successful, and one concludes with Millians that concepts do not have senses,[12] one is left with the task of explaining question (2); what determines the reference of a concept, if not some sort of mode of presentation? According to one view, the reference of concepts is determined through causal chains. This kind of view originated with Kripke's (1980) causal theory of reference: According to this view, the reference of a name is established through an initial baptism and then becomes a rigid

[11] There are further problems raised for descriptivism. Consider for instance this case, also due to Kripke: Tim knows Kurt Gödel only as the person who proved the incompleteness of arithmetic. If descriptivism is true, Tim refers to Gödel via this description. But suppose Gödel was not really the one who proved the incompleteness of arithmetic, but that he stole the proof from his friend Schmidt. In this case, since Schmidt – and not Gödel – is the one who satisfies the associated description, it seems that Tim's concept *Gödel* actually refers to Schmidt. But this result is counterintuitive, and hence poses problems for descriptivists (Kripke 1980). There are further problems posed for descriptivism (see Kallestrup 2012, ch. 2 for an overview), but I will not go into more detail about this. The purpose of the current presentation is simply to point to some of the central problems confronting classical theories of concepts.

[12] 'Sense' understood as a set of associated descriptions. The viability of a Fregean approach (via modes of presentation) does not stand or fall with the viability of descriptivism, since it is possible to hold that modes of presentation is not to be understood as associated descriptions (see Burge 2005, 41. For an account of modes of presentation understood in a non-semantic way see Recanati 2012).

designator. The initial baptism typically involves giving a particular name to a phenomenon, typically accompanied by gestures such as pointing to the relevant object. For instance, Aristotle acquired his name through such a baptism, and this explains why he is the referent of *Aristotle*. When we use the concept *Aristotle* today, we successfully pick out the same individual as did the people being present at his baptism, despite not having been at Aristotle's baptism ourselves. Following Kripke, this is to be explained by us standing in an appropriate causal chain to Aristotle and the people witnessing his baptism. We intend to use *Aristotle* the same way as people used the concept before us. This act of intending to use a term the same way as others is often called *deference*. On this view, when I use a concept, the content of this concept is determined through previous uses of the concept by the people I'm deferring to. Kripke says the following: "In general our reference depends not just on what we think ourselves, but on other people in the community, the history of how the name reached one, and things like that. It is by following such a history that one gets to the reference" (Kripke 1980, 211). Philosophers holding a causal theory of reference about concepts think an analogous mechanism accounts for the content of such entities. A consequence of this view is that one can use concepts without having any descriptions associated with them, as long as one is deferring to others in one's language community. Since, on this picture, the content of concepts are determined by public use, rather than a thinker's own associated descriptions, features external to thinkers are essential for ascribing the appropriate content of concepts and thought to them.[13]

A problem for causal theories of reference is that concepts sometimes seem to change reference. According to Gareth Evans (1973), it seems plausible that *Madagascar* is an instance of a concept changing its reference: Marco Polo intended to use *Madagascar* the same way the locals in Mogadishu used the concept.[14] The locals used the concept to pick out the town on the mainland. Polo, however, made an error and thought the concept was used to pick out the island we now know by that name. Polo's use became standardized, and today we use the concept to refer to the island Polo had in mind. Hence, there seems to be a change in reference of the concept *Madagascar*. This change in reference was due to Polo making a mistake about the referent of an already existing concept. However, Polo *intended* to use the concept the same way as the locals did. If intending to use a concept the same way others in a

[13] Views according to which mental content does not depend solely on intrinsic features of individuals are often labelled *externalist* theories about mental content. I'll return to the notion of externalism in 3.2.
[14] Note that Evans' original account is a criticism of Kripke, and thus formulated with respect to language rather than thoughts.

language community use it ensures the conservation of reference, how can an individual making an error cause a change in reference? The problem posed for causal theories of reference is this: There seems to be a change in reference despite Polo intending to use the concept the same way the locals did, but on a causal theory of reference this should not be possible.[15]

A further problem for the causal theory of reference is that it cannot by itself explain Frege's initial worries regarding theories holding that the relation between syntactic features of a concept and its referent is direct. The difference between the thought *Hesperus is Hesperus* and *Hesperus is Phosphorus* that accounts for the thoughts playing cognitive roles remains unexplained. As already mentioned, if one cannot explain how *Hesperus* and *Phosphorus* can play different roles in cognition – and hence why we are rational in making inferences after learning that Hesperus is Phosphorus that we were not justified in making before the discovery – one fails to account for one of the main explanatory tasks of thoughts; the explanation of rationality. One is then left with two views that both face serious problems. Is it possible to offer a theory that tackles both sets of problems? Originalism, which I present in the next chapter, purports to be just such a theory.[16]

1.2 Chapter Summary

A central question within philosophy of mind concerns the nature of thoughts and concepts. Philosophers adopting a Millian framework for concepts hold that the content of such entities just is their referent; the relation between syntactic features of a concept and its referent is direct. This theory faces problems when explaining how co-referential concepts can play different roles in cognition. In order to explain this, descriptivists introduce a second layer of

[15] For an alternative version of the causal theory of reference, see Devitt (1981). See McKay (1984) for criticism of Devitt's account.

[16] This presentation of the debate between descriptivism and direct reference theories is, of course, simplified. I have focused on the most famous theories within this debate, but there are various different views within each general camp. Consider, for instance the hidden indexical theory of Schiffer (1992): According to this theory it is possible to make justice to the Fregean data within a Millian framework. He thinks there is a (semantic) mode of presentation in addition to direct reference. Originalists (along with Recanati's theory of mental files, which I will present in chapter 7) also set out to explain Fregean data within a Millian framework, but as we shall see, they hold that this can be done without appeal to semantics beyond reference. There are also philosophers who advocate updated versions of descriptivism (See for instance David Chalmers' theory of two-dimensional semantics (e.g. his 2004 and 2006)). I shall not consider these alternatives in this thesis.

semantics to the content of concepts: in addition to having reference, concepts also have a mode of presentation. The mode of presentation is, according to descriptivists, associated descriptions. On this view, the mode of presentation determines the reference in that a concept refers to whatever object satisfies the set of associated descriptions. A problem for this view is that thoughts that share modes of presentation may differ in semantic value; one might be true while the other is false. If rejecting the view that concepts have modes of presentations, one is left with the task of explaining why concepts have their specific reference. On a causal theory of reference the reference of concepts is fixed by an initial baptism, and is then maintained through deference to earlier uses. A problem then is that some concepts, such as *Madagascar* seem to have had a change in reference. How can a change happen despite every user intending to use the concept the same way others use it, if deference ensures the preservation of reference? Further, the causal theory of reference does not explain how co-referential concepts can play different cognitive roles, which was the original problem posed for Millianism.

It seems that we are left with several problems regarding the nature of concepts and thoughts. How should these problems be solved? Sainsbury and Tye's originalist theory provides possible solutions to these and other puzzles rising from this debate. Originalism combines features of the classical theories outlined, but introduces and stresses the importance of the notion of *vehicles* of content to the solutions to the puzzles, so that the result is something quite new. The theories presented in this chapter take the content of concepts to be essential in understanding the nature of concepts; Millians take reference to be essential for the type individuation of concepts while descriptivists hold that concepts are to be typed by way of associated descriptions (to Frege a thought just is sense). Originalists reject the view that semantic features are essential for the type individuation of concepts and thoughts. Instead, they hold that concepts are to be understood as vehicles of content, and must be individuated by way of the origins of their syntactic features. In the next chapter I will give a general introduction to originalism as stated in *Seven Puzzles of Thought,* and point to differences and similarities between originalism and the classical theories presented in this chapter.

Chapter 2

Originalism: A General Introduction to the Theory

In chapter 1 we saw that classical theories of how to individuate concepts and thoughts hold that such entities are to be typed by their semantic properties: A concept or a thought is the concept or thought that it is in virtue of its content. Originalism contrasts with such theories in holding that the essential property of concepts is their historical origin – a property that is not semantic. Concepts are non-eternal abstract objects: They come into existence at particular points in history and they may go out of existence at later points. The point at which a concept comes into existence is the originating use of that concept. One of the key originalist claims is that for every concept there is just one originating use and that every originating use of a concept is the origin of one concept only (Sainsbury & Tye 2011, 3). Originalists hold that concepts are vehicles of content and that these are to be individuated in terms of their origins. While being a completely new theory of how to individuate concepts, originalism as advocated by Sainsbury and Tye adopt many features of classical theories. They use a Millian framework to account for the content of concepts, according to which the content of a concept just is their reference. They agree with Frege that classical Millianism is not sufficient for explaining cognitive significance, but do not agree that one need to introduce further layers of semantics in order to explain the Fregean data. Instead, they hold that the syntactic features of concepts can perform the task of explaining cognitive significance. They agree with Kripke that a concept has a given content in virtue of deference, while agreeing with Evans that concepts may sometimes change reference through time. In this chapter I will give an

overview of originalism as presented in *Seven Puzzles of Thought*; the originalist view on how to individuate concepts and thoughts; how to understand the relation between these two entities; what it takes for a use of a concept to be an originating use; and how concepts and thoughts relate to the contents they express.

2.1 Concepts

In originalist terminology, concepts are vehicles of content that express representational content. On this view concepts have contents, but are not themselves contents: "Concepts are vehicles of representation, tools for thinking" (Sainsbury and Tye 2011, 1). Hence, when giving an account of how to individuate concepts, originalism proposes a way to individuate *vehicles* of content and not the *content* expressed by such entities. This contrasts with the terminology of classical theories that usually take concepts to, at least partly, consist of semantic content.[17] The originalist framework for concepts is modeled on words, which seem to be individuated in terms of their historical properties, rather than their meaning. Analogously, originalists take concepts to be individuated in terms of their origin rather than their content. The central originalist claim is that two concept tokens are of the same type if and only if they have the same origin: "Concept C1 = concept C2 iff the originating use of C1 = the originating use of C2" (Ibid., 4). Since concepts are to be individuated by their origins alone, it may be that two concept tokens that are semantically or epistemically the same may nonetheless be distinct concepts if they have distinct origins.[18] Likewise, two concept tokens that are semantically or epistemically distinct are of the same type if they have the same origin.

One of the main motivations behind originalism is to explain Fregean data within a Millian framework. As we saw in chapter 1, Fregean data is the observation that identity statements may be informative. It's one thing to think that Hesperus is Hesperus and another to think that Hesperus is Phosphorus; one thing to think that Hesperus is visible, another to think that Phosphorus is visible. We saw that classical Millianists, according to whom the meaning of a

[17] Henceforth I will use 'concept', 'vehicle of content' and 'vehicle' interchangeably when discussing originalism.

[18] Concepts that are semantically the same have the same content. Concepts that are epistemically the same appear to the thinker, in some way or another, to be the same. Originalists hold that the content of thoughts and the type of concept a thinker takes concepts to be, is not essential for the individuation of concepts. I return to this later on.

concept just is its reference, cannot explain this data, since *Hesperus* and *Phosphorus* pick out the same object. Sainsbury and Tye agree that we need something more than a theory of direct reference in order to explain Fregean data, but unlike Frege, they do not think that this must be a layer of semantics. Instead, originalists hold that cognitive significance is to be explained in terms of the vehicles of content: "Cognitive processing depends not directly on content but on the vehicles of content: concepts and thoughts" (Sainsbury and Tye 2012, 57). Since cognitive significance depends on vehicles of content rather than representational content, originalists can allow that "distinct thoughts, even if they are referentially isomorphic, can play different cognitive roles" (Sainsbury and Tye 2011, 1-2). The claim that cognitive significance depends on syntactic features of concepts and thoughts is not a new claim (see for instance Fodor (2008)[19]); what *is* a new contribution is that the concepts figuring in thought are to be individuated in terms of their origins. Even though *Hesperus* and *Phosphorus* are referentially isomorphic, they are distinct concepts in virtue of having originated at distinct points in the history (one in the morning, the other at dawn). The concepts being distinct is sufficient for explaining the different roles *Hesperus* and *Phosphorus* play in cognition since originalists allow for the vehicles of content being a separate source of explanation of cognitive significance. I return to this in 3.1.

A further problem for classical Millianism, recall, was that some proper names lack referents. For instance, *Pegasus* does not pick out any object in the world. If the content of a proper name just is its referent, then what should be said about empty concepts such as *Pegasus*? How can we explain the concept *Pegasus* playing an interesting cognitive role if there is nothing more to our theory than a one level Millian view of reference? Also, *Pegasus* and *Vulcan* are both empty concepts, since they lack referents, so how can it be that the two concepts play different roles in cognition? By allowing that cognitive features can be explained by appeal to meaning vehicles rather than content, originalists can give an account of how empty concepts can be cognitively significant: "Some concepts fail to refer, but this does not prevent them having a role in thought" (Sainsbury and Tye 2011, 1). For instance, when Leverrier first introduced the concept *Vulcan* he intended the concept to pick out what

[19] Fodor agrees with the originalist view that Frege cases can be explained without appeal to semantics that go beyond reference. However, the two theories disagree on several matters. One such matter is this: While originalists take concepts to be individuated historically, Fodor thinks that such entities are of different types "when they differ in the (presumably physical) properties to which mental processes are sensitive" (Fodor 2008, 79). On this view, subjects cannot be mistaken about the type and number of concepts deployed in thought. In 3.4 we'll see that originalists disagree; they deny that concepts are transparent to the thinker.

he took to be the planet orbiting between Mercury and the Sun. According to originalists, what is essential for the individuation of the concept *Vulcan* is the point in history at which it was used intentionally for the first time. Since cognitive significance turns on the vehicles of content, according to originalism, what explains the role *Vulcan* plays in thought is the origin of the concept and not the content of the (empty) concept. Whether a concept refers to an object, or fails to do so, is not essential for the cognitive role of the concept, according to originalism. This also explains why *Pegasus* and *Vulcan* play distinct cognitive roles. The concepts, understood simply as vehicles – *solely as syntactic symbols* in the language of thought, have distinct origins and hence they play different roles in cognition. Vehicles seem to play the same role as semantic modes of presentation do in Fregean theories. I return to this in 3.6.

According to Sainsbury and Tye most concepts are public, and hence "concepts are typically sharable" (Sainsbury and Tye 2012, 59).[20] Individuals have their concepts in virtue of being part of a language community, and the participants in the language community share concepts. Young children may come to form their own individual concepts when interacting with the world. For instance they may form a specific concept when interacting with cats. However, the individual concepts children acquire at a young age will typically be replaced by public concepts when the children interact with others in their language community. Sainsbury and Tye take children's willingness to accept correction to be an indication that children replace their individual concepts with public concepts at some point in early development (Ibid., 60). On the originalist account, the concept a child has for picking out cats, introduced independently of other participants in her language community, is distinct from the concept *cat* she use after having acquired the public concept. This is because the individual concept and the public concept were introduced at distinct occasions (I shall discuss the introduction of concepts in a moment). When the child acquires a public concept she will stop using the equivalent individual concept and only use the public concept. It is not up to individuals to decide the nature or content of public concepts: The nature of a public concept is determined by its origin, and the content of such concepts is determined through deference to earlier uses (I return to this in 2.2).

[20] Originalism coincides with Ruth Millikan's theory of concepts in that Millikan agrees with Sainsbury and Tye that Fregean data are to be explained by appeal to sameness and difference in vehicles of content rather than the content expressed by such entities. They also agree that concepts are to be individuated by way of their historical properties. However, while originalists take concepts to be public, Millikan thinks concepts are individual and not sharable: "I have concepts and you have completely other concepts, though many of them may be concepts be of the same thing" (Millikan 2011, 6).

Some concepts used within a language community, however, are not public, according to Sainsbury and Tye. Such concepts include indexical concepts: "It's a feature of indexical concepts that a speaker can introduce them for himself, independently of other thinkers. This contrasts with public concepts acquired by immersion, like the concept *Paderewski*" (Sainsbury and Tye 2012, 52). For instance, your concept *I* is a different concept than my concept *I* since the concepts where introduced at distinct points in history. Your tokens of *I* are of the same type and my tokens of *I* are of the same type, but they are not the same type as each other. Even though our concepts are distinct there is one feature of indexical concepts that is shared amongst participants in a language community; this is what they call a *concept-template*. Such concept-templates are not themselves concepts, but rather rules for forming certain concepts (Ibid., 51). In the case of the concept *I,* the rule given by the content-template is something along the lines of '*I* refers to the person using the concept'. Analogous principles apply to all indexical concept. This explains why individuals within a language community follow the same rules when forming indexical concepts, even though they do not use indexical concepts of the same type. Sainsbury and Tye take public concepts as a starting point and model their theory of indexical concepts on this.[21]

2.2 Originating Use and Deference

The most common way for a concept to come into existence is by way of someone using it intentionally to pick out a phenomenon for the first time. Let me illustrate: In 1963 Murray Gell-Mann introduced the concept *quark.*[22] Before this point in history, no such concept existed.[23] When others in Gell-Mann's community were told about his discovery they also acquired the concept *quark*. When they used the concept they intended to use the same

[21] For someone taking the other direction – taking indexical expressions as a starting point and modelling a theory of lexical expressions on this – see Recanati (2012). I will take a closer look at Recanati's view in chapter 7, where I present his theory as an alterative to Sainsbury and Tye's originalism.

[22] A consequence of this story is that one must allow for the possibility of an object coming before a thinker's mind as an intentional object without the subject already possessing the concepts being originated. Gell-Mann must have been in an intentional relation to quarks before he used the concept *quark* for the first time. I shall not be concerned with the plausibility of this view.

[23] Of course, the same linguistic symbol can be used as a term for distinct concepts. In the case of Gell-Mann, it is know that he borrowed the word *quark* from James Joyce's *Finnegans Wake,* in which Joyce used the word as a term for a different concept than Gell-Mann did. Gell-Mann did not intend his concept *quark* in any way to be the same as Joyce's concept. The origin of our concept *quark* is Gell-Mann's introducing the concept – and not Joyce's – and our concept is therefore to be individuated by this historical event according to originalism.

concept in the same way Gell-Mann did. The intention of using a concept the same way as others in one's language community is what Sainsbury and Tye call deference (Sainsbury and Tye 2012, 70). Originalists take deference to be crucial for a concept's existence through time. When someone uses *quark* today, this is a result of them having accumulated information about the concept from others in their language community. Further, using the concept *quark* today involves deference to earlier uses by the same subject or other people in one's language community. For instance, non-scientists using the concept *quark* intend to use the same concept as the scientists do. The scientists intend to use the same concept, as did scientists before them. This chain of deference goes all the way back to Gell-Mann's introduction of the concept. Gell-Mann, however, did not defer to any other uses of the concept when he first introduced it. Hence, Gell-Mann's first intentional use of the concept *quark* is the origin of that concept. For every atomic concept it is the case that the chain of deference started at some point in past history. The point in history that is the starting point for a certain chain of deference marks the origin of all later concept tokens in that chain. Originalists hold that the origin of the chain of deference a concept token belongs to determines what type it is. All concept tokens that belong to chains of deference with the same origin are of the same type. Now, we need certain conditions for distinguishing between originating uses, which introduce new concepts, and non-originating uses, which simply make use of already existing concepts.

According to Sainsbury and Tye, there are two sufficient conditions for a use of a concept being non-originating:

1) The use involves deference to other uses, by the same subject or other subjects.
2) The use involves informational accumulation from other uses, by the same subject or other subjects (Sainsbury & Tye 2011, 2).

If a concept token belongs to a chain of deference and is not itself the starting point of such a chain, the use is a non-originating use. That is, if an individual intends to use a concept the same way as others in her language community, her use is a non-originating use. According to Sainsbury and Tye, knowledge of the content of concepts is not necessary for someone possessing and using a given concept: "Concept possession is consistent with all sorts of mistakes and misunderstandings about the concept's subject matter" (Sainsbury and Tye

2012, 55).[24] The intention to use the concept in accordance with earlier uses is sufficient for an individual counting as using a given concept. Hence, one cannot use a concept wrongly, according to originalism: "we have no room for a notion of the "correct" use of a concept […] for originalism there is simply the question whether a subject uses or does not use a concept on an occasion. If it is used at all, then it is used "correctly"" (Ibid., 85). This is a causal theory of the vehicles of content, according to which the history of deference is essential for the type individuation of concepts. In 2.4 we shall see that Sainsbury and Tye adopt a similar account of reference.

2.3 Fusion and Fission

A complication for the originalist theory is that some concepts often are taken to have more than one originating use. In the case of the concept *quark*, a standard view is that George Zweig introduced the same concept independently of Gell-Mann. If this is correct, it seems as though the chain of deference of our concept *quark* has two distinct starting points. But, according to originalism, a concept may only have one origin. In order to explain such phenomena Sainsbury and Tye introduce the notion of *conceptual fusion.* In the case of conceptual fusion, two (or more) concepts fuse into one concept. At the time of a conceptual fusion the concepts that fuse together go out of existence and a new concept comes into being. The new concept originates at the point of fusion. This allows Sainsbury and Tye to make sense of the case of Gell-Mann and Zweig within an originalist framework. The concepts introduced by Gell-Mann and Zweig were distinct. At some point, however, Gell-Mann and Zweig's concepts fused into one concept *quark.* The new concept that came out of the fusion

[24] Sainsbury and Tye thus deny Russell's claim that "it is scarcely conceivable that we can make a judgment or entertain a supposition without knowing what it is we are judging or supposing about" (Russell 1912, 58).

is the concept we use today.[25] On this picture we defer to neither Gell-Mann's concept nor Zweig's, but instead we defer to the concept being introduced by the fusion. It is important that "the chain of deference must not run back beyond the point of fusion" (Sainsbury & Tye 2012, 68). This is because the origin of our concept is at the point of fusion. The original concepts give rise to some of the features in the new concept, but the new concept is of a different type than the original concepts.

Another complication for originalism is that two or more concepts often are taken to have the same origin. For instance, the concepts *relativistic mass* and *inertial mass* are often taken to have the same origin, namely Newton's concept *mass*. This story violates key assumptions made by originalism, according to which every originating use of a concept is the origin of one concept only. In order to explain such cases, Sainsbury and Tye introduce the notion of *conceptual fission.* In the case of a conceptual fission, one concept fissions into two (or more) concepts. In this case, the original concept is of a different type than the new concepts that come into being. The new concepts that come out of the fission have their origin in the first intentional use of each of the concepts introduced by the fission. This allows Sainsbury and Tye to give an originalist account of distinct concepts that are standardly taken to have the same origin, such as *relativistic mass* and *inertial mass*: "Features of the concepts *relativistic mass* and *inertial mass* were shaped by the predecessor undifferentiated concept *mass*; but there are three concepts in this story, and three originating uses" (Sainsbury & Tye 2012, 67). Thus, when using the concepts *relativistic mass* and *inertial mass* the chains of deference do not go all the way back Newton's introducing the concept *mass*. Instead, the two chains of deference have distinct starting points: In the case of *relativistic mass* the chain started when someone used this very concept intentionally for the first time, while the deferential chain of

[25] Sainsbury and Tye use the Gell-Mann/Zweig case to illustrate their notion of fusion of concepts. They do, however, note that it might be more historically correct that our concept *quark* today goes directly back to Gell-Mann (Sainsbury and Tye 2012, 68). Even though Zweig came up with a similar concept (which he named "aces") it was Gell-Mann's concept that won through. If this is the case the origin of our concept *quark* is not a case of fusion between Gell-Mann and Zweig's concepts, but instead it goes all the way back to Gell-Mann's first intentional use of the concept. Please note here that the originalist individuation of concepts depends upon historical events, knowledge of which we might not have (and in most cases seem not to have).

inertial mass started when someone used that concept intentionally for the first time.[26]

Sainsbury and Tye don't give a detailed account of fusion and fission, but since in the case of both fusion and fission new concepts come into existence, it is clear that they must be kinds of originating uses. This means that fusion and fission of concepts cannot involve deference to earlier uses, since deference is sufficient for a use being non-originating, according to Sainsbury and Tye. This will be relevant for the discussion in chapter 6, where I propose possible solutions to my thought experiment on behalf of the originalists.

2.4 Content and Deference

According to originalism, concepts are just meaning vehicles. Concepts do, however, *express* contents. As noted earlier, the question about the content of concepts can be divided into two: (1) What is the content of a concept and (2) in virtue of what does a concept come to have the content it has. Originalism in its simplest form is not a theory of how to individuate content, but rather a theory of how to individuate the meaning vehicle expressing such contents. Even so, Sainsbury and Tye's specific formulation of originalism provide an answer to both of these questions. I have already addressed the answer given to the first question: In *Seven Puzzles of Thought* Millianism, which initially was a theory of language and proper names, is transposed to apply to thoughts and atomic concepts.[27] Sainsbury and Tye hold that the content of atomic concepts just is their reference. Atomic concepts that agree in reference agree in content. When giving an account of what it is that makes a concept have a certain content Sainsbury and Tye adopt a causal theory of reference similar to that of Kripke in *Naming and Necessity* (1980). On a simple originalist account of content, concepts acquire their content at their origin, and then maintain that content through time: "the reference of a

[26] Sainsbury and Tye hold that in the case of conceptual fusion the original concept go out of existence: "one can [...] describe the fusion of *a* and *b* into *c* as involving three distinct things, two of which (*a* and *b*) go out of existence as the third comes into existence" Sainsbury and Tye 2012, 68). However, it seems plausible that the original concepts don't necessarily cease to exist. In general, when we use the concept *quark* today, we use the concept introduced by the fusion. However, when thinking about the concepts used by Gell-Mann and Zweig, we use the original concepts, since we intend to use the concepts the same way they did. An analogous comment can be made about the case of conceptual fission: We can use the original concept after the point of fission if we defer to uses of that concept. I don't take this to be a serious objection to the theory: I see no reason why Sainsbury and Tye could not agree with this.

[27] Originalism, does, however, differ from classic Millianism in that originalists hold that there exists no such thing as propositions. According to originalism, thinking doesn't involve standing in a relation to any kind of content. I return to this in 2.5.

concept is *fixed* at its origin and then *preserved* by the same mechanism that preserve the identity of the concept" (Sainsbury & Tye 2012, 69). This mechanism is deference. When someone intends to use a concept the same way as others in her language community, she intends to refer to the same phenomena as they do. This ensures sameness in use amongst all participants in a language community; individuals do not settle anything about the nature or semantic features of public concepts. Hence, reference is to be explained in terms of the history of use rather than in terms of descriptions associated with a given concept.

However, as we saw in chapter 1, this story is not always sufficient; concepts sometimes seem to change their reference through time. Sainsbury and Tye agree with Evans' (1973) observation and say that a concept may change its reference and still remain the same concept. This explains how two concept tokens that are semantically or epistemically distinct may nonetheless be of the same type if they have the same origin. Sainsbury and Tye take the history of the concept *meat* to be an instance of a concept staying the same while the content changes: Originally *meat* referred to anything edible. Due to a slow and gradual drift, the content of *meat* changed. Although everybody who used the concept *meat* deferred to earlier uses, small unnoticeable errors regarding the reference of the concept was made at several occasions during the history. Today we no longer use *meat* to pick out anything edible, but instead we use the concept to pick out animal flesh only. Instead of saying that our concept *meat* is a different concept than the one originally introduced by that word, Sainsbury and Tye hold that our concept and the original concept are of the same type, but with different contents. This, they say, is because "it is a case of gradual drift, with no event that seems a good candidate for the introduction of a new concept" (Sainsbury & Tye 2012, 46). Hence, the content of a concept may change radically through time and still remain the same concept if the development happens gradually and without any intentional deviation of standard use. Since, in the case of *meat*, the concept remains the same it is an instance of neither fission nor fusion; such mechanisms, recall, essentially involve new concepts coming into existence. Their portrayal of the case of *meat* as an instance of a concept changing its content will be important for my thought experiment in chapter 5.

In order to illustrate the difference between a change in concept and a change in the conceptual content, Sainsbury and Tye compare the history of the concept *meat* with a possible history of the concept *Madagascar* - the latter being an instance of change in concepts. The story goes as follows:

> When Marco Polo visited Mogadishu he wrote about the town in his notebook, using the term ”Madeigascar”. Even thought Mogadishu is on the African mainland, Polo described Madeigascar as being an island of great wealth. When some map-makers later came upon Polo’s notebook they gave the name ”Madagascar” to the island we now know by that name, on the basis that Polo seemed clear about it being an island, and that there was only one good candidate for being the island they thought he had had in mind.

Sainsbury and Tye think that this illustrates a change in concepts: “In this case, it seems best to treat either Marco Polo or the map-makers as having introduced a new concept with Madagascar as its referent” (Sainsbury & Tye 2012, 71). The case of Madagascar differs from the history of the concept *meat* in that Marco Polo did not try to acquire or preserve the reference of an existing concept. There is no such case of manifest discontinuity in the history of *meat*, and thus it is not to be treated as a change of concept, according to Sainsbury and Tye. To make the distinction even clearer, one can portray the historical development of the concept *Madagascar* in a different way:

> When visiting Mogadishu, Marco Polo heard some locals use the word “Madagascar” and was thereby introduced to the concept. Polo wanted to use the concept the same way the locals did; when using the concept he intended to defer to the locals’ use. However, Polo made a mistake regarding the referent of the concepts: While the locals used it to refer to Mogadishu, the town on the mainland, Polo thought they used it to refer to the island we now know by the name “Madagascar”. Polo’s mistaken use later became standardized and the locals’ use was forgotten.

Sainsbury and Tye suggest that this alternative story of how our concepts *Madagascar* originated is to be understood as a change in reference rather than a change in concept. The crucial difference between the first and the latter story of how *Madagascar* came into being is that in the latter case Polo, although making an error regarding the concept’s referent, intended to defer to earlier uses of the concept, whereas in the first case neither Polo nor the map-makers intended to acquire and preserve an already existing concept. Note that the latter portrayal of the case of *Madagascar* differs from the history of *meat* in that the story of *Madagascar* does not involve a smooth history where different individuals through the history made several small mistakes about the reference. In the story of *Madagascar,* Marco Polo was the only one who made a reference-changing mistake, but the concept still remained the same, according to originalism. When explaining the change of reference of *meat* Sainsbury and Tye

explicitly say that the concept's remaining the same is due to the smooth history of change in reference, but in the case of *Madagascar* there is no such smooth history. Hence, despite of first appearance, it seems that a gradual drift is not necessary for a change in content of concepts in general, on Sainsbury and Tye's account.

Let me make a few critical remarks about this, since this will be relevant for later discussions. It is striking that Sainsbury and Tye agree with Evans' observations about the change in reference of concepts but maintain a causal theory of reference without explaining how this can be the case. Sainsbury and Tye agree with Kripke that the content of concepts are fixed at their origins and then maintained through time through deference. They hold that the content of concepts essentially depend on the user's intention to use a concept the same way others do. However, the very problem posed by Evans is that in the case of Madagascar the reference changes *despite* Polo intending to use it the same way as others did before him. The fact that Polo successfully uses the same concept as the locals should ensure the sameness in content of Polo's use and the use of the locals to whom he defers, but somehow it fails to do so. If the content of a concept is determined by the public use, as Sainsbury and Tye would have it, how can the content change when a user defers to the public use? Sainsbury and Tye hold that "deference can be modeled (rather over-intellectually) as the recognition that others already use a concept, together with the desire to use the very concept they use, with the very reference it has in their uses" (Sainsbury and Tye 2012, 70). Here it seems that deference is a single desire; that is, it seems not to be the case that someone has two separate desires; one to use the same concept vehicle as others and one to use the concept to express the same content. However, since allowing that a concept may change its reference, originalists must have what one could call a two-dimensional view on deference. The type of the vehicles of content is preserved through deference and the content expressed by concepts is also preserved through deference, but these different aspects of deference can come apart; one may succeed in using the same concept as others and at the same time fail to express the same content expressed by other uses of the same concept. This is what happened in the case of *meat*. But how can the originalist account for the deference being successful at one level and fail at the other?

Might it be that while intending to use the same concept as others, people use the concepts slightly different due to having different *conceptions* of the subject matter? Sainsbury and Tye would disagree; they hold that concepts must be distinguished from conceptions. A conception resembles Fregean sense, understood as a set of associated descriptions: "Call a

"conception" of X a collection of significant beliefs concerning X" (Sainsbury and Tye 2012, 67). According to Sainsbury and Tye, conceptions determine neither concepts nor the representational content expressed by concepts. That is to say, originalists recognize that individuals may conceive of some subject matter in different ways, but they hold that this is irrelevant to thinking. On this view, whenever someone uses a certain concept, the type and content is decided by the public use of that concept regardless of the user's own conception. Whenever someone used the concept *meat* through its history, they intended to use the concept the same way as others did before them. If everyone defers to earlier users of the concept, and the chain of deference goes all the way back to the point in history at which *meat* was first used to pick out anything edible, how can the representational content change if the representational content is determined by deference exclusively?

To illustrate further, consider the following case addressed by Sainsbury and Tye: A philosophy student, Rachel, overhears a conversation at her local café. One person utters the sentence "John Locke was shot". Being an enthusiastic follower of the TV show "Lost", Rachel says to herself "yes, John Locke was indeed shot" because she takes *John Locke* to refer to the fictional character. As it happens, the conversation was really about the character in the TV program, so Rachel's utterance is true. Later on, when at the university, she hears someone else utter "John Locke was shot". Due to her current location, Rachel takes the statement to be about the philosopher John Locke. Knowing that the philosopher John Locke was not shot, Rachel mutters to herself "no, John Locke was certainly not shot". As it happens, the people at the university were also talking about the fictional character named John Locke. Even though Rachel intended the two tokens of *John Locke* to refer to distinct individuals, Sainsbury and Tye take her statements to be about the same individual, due to her deferring to the other people's uses. Her last sentence, then, would be false, and also a contradiction of her first utterance: "In these circumstances, the two utterances (sentence tokens uttered) are contradictory; the one denies what the other asserts" (Sainsbury and Tye 2012, 99). The reason for her sentences being contradictory even though she intended them to be about distinct individuals, is that in both cases she defers to the other users when using the concept *John Locke.* Since the people being deferred to were referring to the same individual, so was Rachel. Hence, the content of Rachel's thoughts are not individuated by way of her own conceptions or what she takes to be the referents of her concepts, but rather by way of the public use of each of the constituent concepts, according to Sainsbury and Tye. Now, let's say one of the small deviations in the history of *meat* consisted in someone using the concept

so as to pick out anything edible but white sesame seeds. That is, if asked whether white sesame seeds are in fact meat, this individual would respond negatively. Since the mechanism of deference ensured that Rachel's concept token *John Locke* referred to the fictional character (even if she thought she was referring to John Locke the philosopher), we should expect the content of someone wrongly taking *meat* to pick out anything edible but white sesame seeds to include white sesame seeds since white sesame seeds, at the time, were part of the public content of *meat*. It seems then that it shouldn't be possible for a concept to have a change in reference on Sainsbury and Tye's view, but even so they allow for this.

As noted, Sainsbury and Tye provide no explanation of how the change in reference is supposed to take place, and there seems to be no obvious answer to give on behalf of the originalist. I return to this point in chapter 5, where I argue that Sainsbury and Tye's account fails to explain certain cases of rationality, and hence that it fails as a general theory of concepts and thoughts. Let us now turn to the originalist theory of thoughts.

2.5 Thoughts

The notion of concepts that is relevant to originalism is one according to which concepts are representational constituents of thoughts. A thought is a well-formed structure of concepts. Hence, the originalist theory of concepts also provides a new way of individuating thoughts. A thought is to be individuated by way of its constituent concepts; the individuation of thoughts depends on the origins of the constituent concepts. Two thought tokens are of the same type if and only if all the concepts figuring in one also figures in the same structure in the other. It is clear that originalism endorses a form of compositionalism: the type of thoughts is determined by the constituent concepts and how these are structured. Originalism is, however, incompatible with traditional compositionalism, according to which the *content* of the constituent concepts determines the content of more complex structures such as thoughts. The reason why originalists must abandon classical compositionalism is that classical compositionalism leads to an unrestricted substitution principle, the principle that concepts that agree in content can be substituted without restriction. For instance, the belief *the thought that Greeks are Greeks is exactly as informative as the thought that Greeks are Greeks* is true. However, the belief *the thought that Greeks are Greeks is exactly as informative as the thought that Greeks are Hellenes* is not true. But, according to an

unrestricted substitutional principle, the two beliefs agree in content, and one should therefore be able to substitute one with the other (Sainsbury and Tye 2012, 74). Sainsbury and Tye hold that the content of a thought is a "(possibly empty) set of possible worlds, namely the set in which the relevant conceptual structure is true" (Ibid., 111). Thoughts that share a set of possible worlds in which they are true, agree in content. Thoughts that do not share such a set of possible worlds differ in their representational content. Since the two beliefs about the informativeness of thoughts involving the concepts *Greeks* and *Hellenes* do not share a set of possible worlds in which they are true, they have distinct contents, according to Sainsbury and Tye. Hence, originalism is incompatible with an unrestricted substitutional principle, and must therefore abandon traditional compositionalism. Nonetheless, compositionality is maintained at the level of vehicles.

The content expressed by a thought is not, according to originalism, essential for a thought being a certain type. In fact, originalists hold that thinking does not involve standing in a relation to any kind of content: "we deny that thinking (believing, etc.) consists in bearing an appropriate psychological relation to *any* sort of content. […] Our position is that the content of the thought is not to be identified with what is thought" (Sainsbury and Tye 2012, 110). That is, thinkers do not stand in a direct relation to the content of thoughts (sets of possible worlds) but rather they "relate to sets of worlds via thoughts" (Sainsbury and Tye 2011, 115). This involves giving up the traditional idea that thinking involves standing in a relation to a proposition (be it Fregean – that is, consisting of modes of presentations – or Russellian – that is, consisting of objects – in nature). Rather, thinking involves standing in an appropriate psychological relation to a conceptual structure (meaning vehicles, that is), according to Sainsbury and Tye.[28] This sets them apart from similar theories, such as Fodor's (1975, 2008) according to which thinking involves the thought token (individuated functionally) standing in a semantically appropriate relation to a proposition. To illustrate: On Fodor's view, for a thinker to believe that penguins are birds involves a triadic relation between the thinker, a mental representation in her head (a sentence in the language of thought composed of the syntactic elements *penguin* and *bird* arranged in a certain way), and the proposition that

[28] Sainsbury and Tye agree with David Lewis (1986) that the content of thoughts are sets of possible worlds. The theories differ, however, on how to characterize such contents: Lewis takes the content of thoughts to be propositions, for on his account propositions simply are sets of possible worlds, whereas Sainsbury and Tye do not characterize the content of thoughts as propositions, since they do not seem to share Lewis' view. This is clear from the fact that they deny that thinking involves standing in a relation to propositions (propositional content is traditionally taken to be what one stands in a relation to when thinking). It is not clear how Sainsbury and Tye's account differs from Lewis' – perhaps it is a terminological dispute. I shall not go into this.

penguins are birds. On the originalist view, however, for a thinker to believe that penguins are birds involves only a relation between the thinker and the vehicle (which is a structure consisting of the concepts *penguin* and *bird* arranged in a certain way). The vehicle expresses a representational content – which is not to be understood as the proposition that penguins are birds, but as the set of worlds in which the thought (i.e. vehicle) *penguins are birds* is true – but the thinker does not stand in a direct relation to such a content.

Just as concepts can be of the same type even if semantically distinct, two thoughts of the same type may also differ in content: "Given that thoughts are structures of concepts, allowing that a concept can change its reference entails allowing that a thought can change its truth condition" (Sainsbury & Tye 2012, 72). Sainsbury and Tye adopt what they call a minimal representationalist framework, according to which thoughts that differ in truth conditions differ in representational content. For instance, the first users of the concept *meat* may have formed thoughts like *squash is meat*. This thought would be true at the time when *meat* referred to anything edible. If someone was to form the same thought today, given that the concept *meat* remains the same, her thought would be structurally and conceptually the same as that of the early users, but her thought would be false, since today *meat* only picks out animal flesh. Since the two thought tokens differ in truth value they have different representational content. According to originalism, then, two thought tokens may be of the same type even if they differ in representational content.

Further, it is not just the type of thoughts that is to be determined by way of their constituent concepts; the consistency of thoughts also depends directly on the concepts figuring in the thoughts and how these are structured: "Inconsistent thoughts are contradictory iff one consists of the other embedded in a concept for negation. If one thought contains a nominal concept, a contradiction must contain the same nominal concept at the corresponding position in the structure" (Sainsbury & Tye 2012, 135). Consider for instance the thoughts *the cat is on the mat* and *the cat is not on the mat.* Since the two thoughts share all concepts structured the same way, except from the latter containing a negation of one of the nominal concepts in the former, the two thoughts are contradictory. Notice that this explanation of the consistency of thoughts appeals only to the vehicles of content and how they are structured, and not the content expressed by such entities. I return to this feature of Sainsbury and Tye's theory in chapter 5 where I argue that the theory fails to explain rationality.

2.6 Chapter Summary

In chapter 1 we saw that there are many problems concerning classic theories of concepts and thoughts. Sainsbury and Tye combine what they take to be the best features of these theories in order to avoid the problems posed for traditional direct reference theories, descriptivism and causal theories of reference. As we have seen in this chapter, originalists adopt a Millian framework for the content of concepts, but they agree with Frege that something more than classical Millianism is needed in order to explain cognitive significance. They do not agree with Frege, however, that what is needed is a second layer of semantics. Instead they hold that cognitive significance turns on the *syntactic* features of concepts; the vehicles of content. More specifically, they hold that the origins of concepts is essential for their type individuation and that this explains their role in cognition.

Let me here restate the central claims of originalism, as they will be important in later chapters, where I take issue with several of them. The key claim of originalism is that two concept tokens are of the same type if and only if they have the same origin. Concepts originate when someone uses them intentionally for the first time. There are two sufficient conditions for a use being a non-originating one: the use involves deference to earlier uses and the use involves information accumulation from other users. A further claim essential to originalism is that every concept has only one origin and that the origin of a concept is the origin of one concept only. A complication for the theory is that some concepts often are taken to have more than one origin and also that more than one concept have the same origin. Originalists explain this by appeal to the notion of conceptual fusion and fission. In the case of conceptual fusion and fission, new concepts come into being. These concepts bear some similarity to the original concepts but are of distinct types. Since conceptual fusion and fission involves new concepts originating, they must be due to an intentional introduction and not involve any deference to earlier uses.

Originalism is a compositional theory of thoughts: A thought is a well-formed structure of concepts. The contents of the concepts deployed in a thought combine to generate the representational content of that thought (but as I noted, they deny an unrestricted substitutional principle). The content of thoughts are the set of world in which they are true. Thoughts are, however, not to be type individuated in terms of their representational content. Instead, the individuation of thoughts is to be by way of it syntactic features; the concepts figuring in the thought and how these are structured. Two thought tokens are of the same type

if and only if all the concepts figuring in one also figures in the same structure in the other. Since concepts are to be individuated by way of their origins, on this view, the individuation of thoughts depend directly on the origin of the constitutive concepts. Further, originalists hold that the cognitive significance of thoughts depends on the type of concepts figuring in thought and not directly on the content expressed by such entities. This is why identity statements can be informative; if the concepts figuring in the thoughts have distinct origins, the mental processing of these concepts is distinct.

Sainsbury and Tye subscribe to a causal theory of reference c.f. Kripke, but agree with Evans that concepts may change referents while remaining the same concepts. I have pointed out that this is problematic, since they do not explain how someone can successfully use a public concept through deference, but use it to pick out something else than what is picked out by the public concept. If a concept has a certain content in virtue of the mechanism of deference, one should expect the content to be stable. In the case of *meat*, however, the concept remained the same but the content changed gradually through time, according to Sainsbury and Tye. If deference ensures sameness in use, such change in referent seems implausible. This was the problem Evans posed for Kripke, and Sainsbury and Tye do not give any explanation as to how this can be the case. The view that concepts may change reference combined with the claim that cognitive significance turns on the origins of concepts is what makes the basis for my criticism of originalism in chapter 5. I turn now to the main case in favour of originalism: its ability to solve central puzzles of thought.

Chapter 3

The Seven Puzzles of Thought and The Originalist Solutions

In philosophical theorising about thoughts and concepts, several central problem cases or puzzles arise against which one can test a given theory. A failure to solve such central problems casts doubt on a theory, whilst its ability to solve them counts in its favour. Originalism is proposed in order to solve seven central puzzles that have occupied philosophers of mind over the past century. Having presented originalism in Chapter 2, I now turn to its solutions to these puzzles, which is the main consideration in support of the theory. If it can in fact solve the puzzles, we have good reason to believe it to be true. As we have seen, Sainsbury and Tye hold that one does not need a sophisticated account of mental content in order to solve the puzzles: they take the solutions to the puzzles to depend on the vehicles of content rather than the semantic features of thoughts[29]. We shall now see how this works. I present the seven puzzles along with the originalist solutions in turn, along with minor comments. Though I shall foreshadow certain criticisms, I leave the main discussion and criticism of these solutions to chapter 4, where I seek to demonstrate that the theory fails to account for three of the puzzles, and thus cast doubt on the theory. Considering their solutions to the puzzles also serves to make clearer certain commitments of the theory: these I shall note along the way.

[29] Though for this reason, as we shall see in chapter 4, it is not clear that some of the solutions they provide actually engage with the classical puzzles, as the latter seem to concern *content,* rather than vehicles.

3.1 The Puzzle of Hesperus and Phosphorus

The first puzzle is the puzzle of *Hesperus* and *Phosphorus*, stemming from Frege (1892).[30] The concepts *Hesperus* and *Phosphorus* have the same reference: the planet Venus. On a Millian framework – according to which the content of a concept just is its reference - the representational content of the two concepts should be the same. If all there is to a thought is its reference, it seems that the thought *Hesperus is Hesperus* should have the same cognitive significance as the thought *Hesperus is Phosphorus*. However, the former is trivial whereas the latter is informative; hence the two thoughts do not play the same role in cognition. The difference in informativeness cannot be explained by classical Millianism; we seem to be in need of something more than reference in order to explain the difference in the two thoughts. Frege's proposal, recall, was that the two thoughts differ in content due to a difference in mode of presentation. Sainsbury and Tye, on the other hand, hold that the two concepts do in fact have the same content, since originalists agree with Millians that the content of concepts just is their reference. Unlike Frege, then, Sainsbury and Tye cannot explain the difference in informativeness in terms of content.

The originalist solution to the puzzle is to appeal to a difference in cognitive significance due to a difference in vehicles of content. Originalists hold, as we have noted, that "cognitive processing depends not directly on content but on the vehicles of content: concepts and thoughts" (Sainsbury and Tye 2012, 57). This is because thinkers only stand in relation to the content of thoughts (sets of worlds) *indirectly*, via thoughts. *Hesperus* and *Phosphorus* are distinct concepts, according to originalism, since they originated at numerically distinct events, so the two concepts play different roles in cognition. When forming a thought containing the concept *Hesperus* this involves a distinct relation to the planet Venus from a thought containing the concept *Phosphorus*. The thought *Hesperus is Hesperus* is an identity statement between two tokens of the same concept type. In the thought *Hesperus is Phosphorus* the identity statement is between concept tokens of different types, due to the concepts having distinct origins. This explains why the first thought is trivial while the latter is not, even though they have the same representational content on a Millian framework of semantics. The vehicles of content seem to play the role Frege's modes of presentation did, and yet no further layer of semantics has been introduced. Sainsbury and Tye claim that, since the thought *Hesperus is Hesperus* has the same truth conditions as the thought *Hesperus is*

[30] Recall the discussion of this in chapter 1.

Phosphorus, the Ancient Babylonians discovered no new fact about the world when they learned that Hesperus is Phosphorus. Rather, when the Ancient Babylonians discovered that Hesperus is Phosphorus, they made a *cognitive discovery*: They discovered something new about their concepts, namely that *Hesperus* and *Phosphorus* are co-referential (Ibid., 125).[31]

When discussing the case of Hesperus and Phosphorus, Sainsbury and Tye address a corresponding case: How can it be that the thought *Greeks are Greeks* is trivial, while the thought that *Greeks are Hellenes* is informative, given that *Greeks* and *Hellenes* have the same reference? Sainsbury and Tye appeal to a difference in cognitive processing effort involved in the two thoughts:

> The thought that Greeks are Greeks is typically uninformative, whereas the thought that Greeks are Hellenes is potentially informative. In processing the first, only one concept is exercised, though on two occasions. Whatever processing effort is required has already been made by the time the second occurrence of the concept is encountered; the previous interpretive outcome can simply be brought forward (Sainsbury and Tye 2012, 54).

The explanation of the cognitive significance of informative identity statements in terms of cognitive processing effort indicates that Sainsbury and Tye have a view on cognition resembling that of the computational theories of mind (CTM).[32] According to this view, the mind is structured the same way as modern computers; mental processes are understood as computations involving syntactic symbols. One of the main motivations behind CTM is that one of the main instances of a rational process, deductive reasoning, can be "characterized in terms of relations among *syntactically specified sentences* in a formal language that can receive a *systematic semantic interpretation*" (Rey 1997, 212). The rules of first order valid arguments can be specified in terms of syntactic features alone. The key claim of CTM is that mental processes are computational processes of the syntactic features of thoughts.[33] According to Sainsbury and Tye, then, the reason why the thought *Hesperus is Hesperus* is trivial whereas the thought *Hesperus is Phosphorus* is informative, is that the latter thought requires more cognitive processing effort than the former, due to the number of concepts involved in the thought; that is, due to their syntactic features.

---

[31] C.f. Frege's initial view in his (1879). Frege later rejects this view in his (1892). In 4.1, we shall see that Frege's reasons for rejecting this view is also a reason for rejecting the originalist solution to the puzzle.
[32] See for instance Fodor (1975, 1980 & 2008) and Rey (1997).
[33] Note that the claim that cognitive processing can be defined without appeal to semantics does not entail that concepts and thoughts do not have contents.

According to Sainsbury and Tye, then, the informativeness and triviality of identity statements is to be explained in terms of cognitive processing effort. Originalists agree with CTM that syntactic features of concepts and thoughts are essential for cognitive processing effort. This is because, according to originalism, thinking does not involve standing in a relation to any kind of content (Sainsbury and Tye 2012, 110).[34] The question then, is how these syntactic features are to be individuated. Originalists, of course, take such features to be individuated by their origins. But why should we believe that the origins of the concepts deployed in thought are essential for the way the brain processes thoughts?[35] It seems clear that most of the time the origins of the concepts we use are inaccessible to our cognitive systems (though we may in some cases have knowledge of origins), and, as we'll see in 3.4, originalists are committed to the view that cognitive processing effort sometimes depend on the number of concepts a thinker *takes* a thought to contain rather than the actual number of different concepts deployed in thought. In chapter 5 I will argue that, if originalism is true, it is possible for someone to use two concept tokens of the same type to express contradictory contents. In that case, how are originalists to explain someone being rational in accepting or rejecting simple deductions containing syntactically identical concept tokens, but with contradictory content?[36]

3.2 The Puzzle of Twins

The second puzzle addressed by Sainsbury and Tye is the puzzle of twins. The puzzle of twins stems from two different puzzles introduced by Hilary Putnam (1975) and Tyler Burge (1979b). Putnam offers the following case: Imagine a distant planet that is completely identical to ours, particle by particle. This planet has duplicates of all inhabitants on Earth.

[34] Note that CTM does not in itself entail the view that thinking doesn't involve standing in a relation to any kind of content. Fodor (2008) for instance takes thinking to involve standing in an appropriate relation to propositions.

[35] Cf. Fodor 1980, where it is claimed that syntactic computational processes take into account only the *shapes* of symbols. The point I'm making here is that in light of this traditional understanding, it is not clear how such processing can take into account *origins*, given that these may not be reflected in the shapes of symbols (Sainsbury and Tye do not offer a detailed account of this, but see pp. 86-88 in their (2012)). I shall return to it.

[36] This problem does not seem to arise on traditional CTM accounts, for on such a view, a single sentence in the language of thought cannot potentially express multiple differing contents (there cannot be ambiguity in the language of thought) (see again Fodor 1980, though see Fodor 1994).

Let's call this planet *Twin Earth*.[37] Twin Earth is identical to Earth, except for one thing; on Twin Earth, everything that on Earth has the chemical structure H2O, is made up of molecules with chemical structure XYZ. Every individual on Earth has an intrinsic duplicate on Twin Earth. For instance, on Earth there is a person Oscar and on Twin Earth there is an exact copy of Oscar, namely Twin Oscar. Now, consider Oscar forming the belief *water is wet.* At the same time, on Twin Earth, Twin Oscar also forms the thought *water is wet*. However, while Oscar's thought picks out H2O, Twin Oscar's thought refers to XYZ. Hence, Oscar and Twin Oscar's thoughts have different truth conditions; Oscar's thought is true if water (H2O) is wet, whereas Twin Oscar's thought is true if twin water (XYZ) is wet.[38] A common view is that a difference in reference indicates a difference in mental content. But how can there be a difference in the content of the twins' thoughts if Oscar and Twin Oscar share all intrinsic physical properties? Although the classic debate concerns mental content, Sainsbury and Tye take the puzzle to follow from such cases to be how two intrinsic duplicates can think different thoughts (vehicles of content): "Mental properties are intrinsic, and thoughts are mental, so twins shouldn't be able to think different thoughts!" (Sainsbury and Tye 2012, 9). Their solution to the puzzle is to say that Oscar and Twin Oscar have distinct *water* concepts, since the concepts have distinct origins: Oscar's concept originated at Earth, while Twin Oscar's concept originated on Twin Earth. Since the concepts are distinct, the thoughts as a whole are distinct. The thoughts being distinct due to a difference in syntax explains why they play different roles in cognition, according to Sainsbury and Tye.[39]

As we've seen, originalists hold that thoughts are to be individuated by way of their syntactic structure and the origins of the compositional concepts. Thoughts that are semantically or epistemically the same may nonetheless be of distinct types if they contain distinct concepts. The originalist solution to the puzzle of twins entails that individuals that are intrinsic duplicates may differ with respect to their thoughts. Originalism, then, is committed to *concept externalism.* In *Seven Puzzles of Thought* originalist concept externalism is stated as

[37] The original Twin Earth thought experiment is due to Putnam (1975). My presentation of the puzzle is analogous to the one presented by Putnam, but where the original example was formulated to be about language I will transpose it to be about thoughts.

[38] Note that it is not important whether Oscar and his duplicate are in fact aware of the molecular composition of water: According to Putnam, the conclusion holds even though we imagine the time of the utterances to be in 1750 – before any knowledge of the molecular structure of water. Oscar and Twin Oscar "understood the term "water" differently in 1750 *although they were in the same psychological state,* and although, given the state of science at the time, it would have taken their scientific communities about fifty years to discover that they understood the term "water" differently" (Putnam 1975, 11).

[39] When presenting the puzzle of twins, Sainsbury and Tye lay out both Putnam's and Burge's thought experiments. When giving a solution to the puzzle, however, they only discuss Putnam's Twin Earth case. I return to this in 4.3.

follows: "If an atomic concept *C* has an origin *O*, then it is metaphysically necessary that *C* has origin *O*" (Sainsbury & Tye 2012, 109). It is metaphysically necessary for a concept to have its specific origin since had it had some other origins it would not be the same concept. That is to say, in a possible world where *quark* has a different origin than it actually has, the concept would be distinct from the actual concept. This shows that intrinsic duplicates may differ with respect to their conceptual repertoires: they may have different vehicles of content.[40]

Note that originalist concept externalism is not the same as *semantic externalism.* The latter is the view that it is metaphysically possible for intrinsic duplicates to differ with respect to their mental content. In contrast, *semantic internalism* is the view that intrinsic duplicates cannot have qualitatively distinct mental contents. Originalist concept externalism is compatible with both internalism and externalism about mental content. Consider again Putnam's Twin Earth case. Originalism provides an answer as to how it is possible for the intrinsic duplicates to differ with respect to their thoughts (i.e. vehicles of content), but it does not give an answer as to how to individuate their mental content. Since the concepts *water* as we use it on Earth and *water* as used on Twin Earth originated in numerical distinct events – one on Earth and the other on Twin Earth – the thought *that is water* entertained by an individual on Earth is distinct from the thought *that is water* entertained by an individual on Twin Earth, but originalism in its simplest form (i.e. without presupposing a Millian understanding of content) is neutral on how the content of such thoughts are to be individuated. That is to say, even though the twins in the thought experiment have thoughts that are distinct in virtue of their concept *water* having distinct origins, originalist concept externalism is compatible with the content of such thoughts being the same; even if the content of the concepts *water* and *twin water* is to be individuated in terms of associated description (e.g. 'clear, drinkable liquid'), and thus share content, Oscar and Twin Oscar's thoughts would still be distinct due to their concepts having distinct origins. The fact that originalism is compatible with both internalism and externalism might seem like a virtue of the theory, but I will show in 4.3 that this actually

[40] This is a further respect it which Sainsbury and Tye's originalism differs from traditional theories that accept the language of thought hypothesis. For on such views, if two persons share all intrinsic physical properties, they necessarily share all computational properties. They must have the same concepts, i.e. vehicles of content. Sainsbury and Tye do not explain the notion of computational properties at work, that can allow for intrinsic properties and computational properties to come apart. This is related to the points in footnotes 35 and 36. It also raises questions about the mind-body problem and mental causation that I will not go into. See again Fodor 1980 and 1994.

renders the originalist solution irrelevant to the classical puzzle of twins; the difference in concepts between intrinsic duplicates does not serve any real explanatory purpose.

Further, although originalist concept externalism does not entail semantic externalism, Sainsbury and Tye's specific version of originalism *is* committed to semantic externalism. This is because they adopt a Millian framework of reference, according to which the representational content of atomic concepts just is their reference. Atomic concepts that agree in reference agree in content, and atomic concepts that have distinct referents have distinct content. Hence, Sainsbury and Tye would say that Oscar and Twin Oscar differ with respect to mental content as well as their thoughts, since Oscar's concept *water* refers to H2O while Twin Oscar's concept *water* refers to XYZ. Further, according to Sainsbury and Tye concepts are public and the use of a concept involves deference to others in one's language community. The content of concepts is not determined by the user's own conceptions (i.e. the set of significant beliefs concerning the referent), but rather by how the concept is used in one's community. Hence, Sainsbury and Tye's account as a whole *is* incompatible with content internalism. Let it be clear that Sainsbury and Tye explicitly agree with this: "Originalism does not entail semantic externalism. Nonetheless, we ourselves accept semantic externalism and many of our remarks in earlier chapters reflect this acceptance or provide reasons for it" (Sainsbury and Tye 2012, 90). Their endorsement of semantic externalism will be relevant for my criticism of Sainsbury and Tye's solution to the puzzle of twins in 4.3.

3.3 The Puzzle of Cats and Chats

The third puzzle – the puzzle of cats and chats – is due to Brian Loar (1987). It goes as follows: Paul belongs to an English speaking community but is brought up by a French nanny. The nanny speaks English with Paul, except when talking about cats; then she uses the French word 'chat'. Paul picks up on his nanny's words and comes to form thoughts such as *all chats have tails*. It seems plausible to hold that Paul has acquired the concept *cat.* Intuitively, since Paul uses the concept the same way his nanny does, when thinking thoughts such as *all chats have tails* he expresses the belief that all cats have tails. One day Paul's parents meet up with Paul at a hotel room in London. During this meeting, Paul does not encounter any depictions of cats or any real cats. However, his parents tell him stories about cats. Paul does not know that cats just are chats. From his parents' stories, Paul comes to know, amongst other things,

that cats have tails. But doesn't Paul already have the belief that all cats have tails? "There are powerful reasons to say that the belief that Paul expresses with "All cats have tails" is the same as the belief that he expresses with "All chats have tails"" (Sainsbury and Tye 2012, 11). One day Paul's nanny decides to tell Paul that cats and chats are the same. When learning this, it seems like Paul makes a discovery, but the thought that cats are chats seems to represent just what the thought that cats are cats represents, so how can there be anything to discover? This puzzle bears similarity to the case of Hesperus and Phosphorus discussed in 3.1, but the puzzles differ in that it seems more problematic to give a Fregean explanation of the puzzle of *cats* and *chats*. In the case of *Hesperus* and *Phosphorus,* it seems reasonable that the two concepts differ in meaning and thus play different roles in cognition as a result of this. In the case of *cat* and *chat,* on the other hand, it seems plausible that the concepts have the same meaning since the word 'chat' translates to 'cat' in English.

Sainsbury and Tye claim that it is plausible that *cats* and *chats* have distinct origins and thus that they are distinct concepts. The thoughts *cats have tails* and *chats have tail,* then, are not of the same type. This explains why the thoughts can play distinct cognitive roles in Paul's mental life, and thus why he makes a discovery when he learns that the distinct concepts *cats* and *chats* have the same representational content. However, Sainsbury and Tye admit that it may turn out that *cats* and *chats* have the same origin and therefore are the same concept. If this is the case, Paul's beliefs do not only express the same content, they are also of the same type. How, then, can there be anything for Paul to discover? For this version of the puzzle, Sainsbury and Tye hold that the explanation must be linguistic. Before making the discovery the difference in spelling and pronunciation makes Paul believe that *cats* and *chats* are distinct concepts. When his nanny tells him that cats are chats, Paul "gains only the metalinguistic knowledge that "cats" refers to what "chats" refers to" (Sainsbury and Tye 2012, 129). In this way Sainsbury and Tye explain how it can be that Paul makes a discovery when learning that cats are chats, even if the concepts have the same origin. I elaborate further on the possibility of individuals being mistaken about the number of concept they possess in 3.4.

3.4 The Puzzle of Paderewski

The fourth puzzle addressed by Sainsbury and Tye is Kripke's (1979) famous puzzle of Paderewski. The puzzle runs as follows: Ignace Paderewski was a popular polish pianist. He

was also engaged in politics and after the First World War he became the Polish prime minister. Now consider a person, Peter, who independently comes to know about Paderewski the piano player and Paderewski the politician without realizing that they are in fact the same person. Instead Peter believes that it is a case of two different people sharing the same name. Having been to one of Paderewski's concerts, Peter forms the belief *Paderewski has musical talent.* However, Peter also believes, on good authority, that no politicians have musical talent and that no pianists are politicians. Thus it also seems plausible to attribute to Peter the belief *Paderewski lacks musical talent.* Peter is justified in entertaining the latter belief having learned from reliable sources that no politician has musical talent. The problem, then, is that it seems that Peter cannot have both of his Paderewski-beliefs at the same time. This is because the beliefs are inconsistent. Which of these beliefs are we to ascribe to Peter? It seems to be the case that Peter really holds both of these beliefs: If asked at a concert whether he thinks Paderewski has musical talent he will give a positive answer, but if asked the same question when attending one of Paderewski's political talks, Peter would give a negative answer. But if Peter holds this set of contradictory beliefs he would be irrational, and *ex hypothesi* he is not. According to originalism there is just one public concept *Paderewski* so his thoughts are contradictory even on an originalist framework:

> According to originalism, there is just one public concept PADEREWSKI, which Peter exercises both when he forms the belief that Paderewski has musical talent, and when he forms the belief that Paderewski lacks musical talent. In the originalist framework, Peter has contradictory beliefs: apart from negation, the beliefs are made up of just the same concepts in the same position. The challenge is to explain how Peter can, nonetheless, be rational. If a rational thinker can believe contradictions, do we not lose all grip on what makes a thinker rational? (Sainsbury and Tye (forthcoming), 1-2).

Peter's beliefs *Paderewski has musical talent* and *Paderewski does not have musical talent* are contradictory at the level of content vehicles, and the originalists must therefore appeal to something else beyond their core framework in order to explain Peter's being rational.

Sainsbury and Tye's solution to the puzzle is to say that Peter has a false belief about what he believes. That is to say, Peter has a false second order belief: He falsely believes that he does not believe that Paderewski has musical talent. In order to illustrate how this can be the case, imagine the following. Two weeks after attending Paderewski's concert Peter goes to see Paderewski give a political talk. If, at the talk, someone were to ask Peter whether he thinks Paderewski has musical talent or not, he would answer that he does not believe this to be the case. Peter is, however, wrong when giving this reply. That is to say, his statement that he

does not believe it to be the case that Paderewski has musical talent is false. This is because when attending the concert two weeks earlier, Peter formed the belief that Paderewski has musical talent, and he has not abandoned that belief by the time of the political talk: "In fact he does believe that Paderewski has musical talent, though he does not believe he believes this" (Sainsbury & Tye 2012, 137). Sainsbury and Tye take Peter having a false second order belief to account for him being rational in holding contradictory first order beliefs.

When solving the puzzle, Sainsbury and Tye propose an account of what they take to be a correct view about rationality; a view according to which individuals can be wrong about the number of concepts they possess. Peter does have the contradictory beliefs *Paderewski has musical talent* and *Paderewski does not have musical talent,* but since he falsely believes that he does not believe that Paderewski has musical talent – he takes himself to have two Paderewski concepts, when he in fact has one – he is still to be considered rational. A common view is that individuals have privileged access to their own conscious mental states, including their thoughts; someone is in a position to know whether his or her own thoughts are of the same type or not. Intuitively, one can know which concepts one possesses and whether they are of the same type or not. The thesis of introspective knowledge of comparative concepts (IKCC) may be stated as follows:

> IKCC: When our faculty of introspection is working normally, we can know apriori via introspection with respect to any two present, occurrent thoughts whether they exercise the same or different concepts (Sainsbury & Tye 2012, 92).

A consequence of originalism is that IKCC must be abandoned. Since, in most cases, the exact point in history at which a given concept was first used intentionally is unknown, individuals do not possess the necessary knowledge to decide which concepts are the same and which are not. This has consequences for cases of *slow switching*. In slow switching thought experiments (originally introduced by Burge (1988)) someone is, without knowing it, teleported from Earth to Twin Earth. For instance, Oscar grows up on Earth and has the concept *water*. One day, unbeknownst to him he is teleported to Twin Earth. Oscar has no way of knowing that he has been switched to a different planet, since Twin Earth is qualitatively identical to Earth (except from water being made up of XYZ, of course). Gradually, by deferring to other people in his new language community on Twin Earth, Peter comes to use the Twin Earth concept *water* when talking about the clear, drinkable liquid in his surroundings. Peter, however, is not aware of this change in concepts, since he is unaware

that he defers to a different group of people's use of the concept *water* than did he as a child. However, Sainsbury and Tye hold that when Oscar thinks back on his early childhood and is forming such beliefs as *I used to drink ten glasses of water a day when I was a child*, his old concept *water* that originated on Earth is active in his mind, so his thought is true. Oscar is not aware of his using distinct water concepts, and thinks that he is using the same concept when talking about his drinking water today and his drinking water as a child. Oscar's thought *the water I drank as a child tasted better than the water I get served nowadays* contains two distinct *water* concepts, but Oscar is in no position to discover this. Further, in the case of Peter in the Paderewski case, Peter does not know that his belief that Paderewski has musical talent contains the same concept *Paderewski* as his belief that Paderewski lacks musical talent.[41] Hence, abandoning IKCC is essential to Sainsbury and Tye's solution to the puzzle of Paderewski. I offer a criticism of Sainsbury and Tye's solution to the Paderewski case in 4.1.

3.5 The Puzzle of the Two Tubes

The fifth puzzle is the puzzle of pure demonstratives. To present the puzzle, we may use a case constructed by David Austin (1990): Jonathan is able to focus his eyes independently of each other. One day Jonathan looks through two tubes that are pointed in different directions but that come together at the other end. At the point where the tubes meet there is a red dot with diameter the same as the tubes. Jonathan does not know how the tubes are oriented, but he sees just one red patch. Jonathan believes that he may be subject to a complex medical condition, the effect of which is that he cannot tell on the basis of perception where objects are located, and further that he cannot know which eye he is using to see which object. When seeing the red dot in each of his eyes, he wonders whether that (referring to the red dot he actually is seeing with his left eye) is identical to that (referring to the red dot he actually is seeing with his right eye). This is not a trivial question: Jonathan genuinely does not know the answer. When thinking about this, Jonathan forms the thought *that is that,* but he does not

[41] If one agrees with originalist concept externalism and externalism about mental content, one must deny that one can know via introspection the content of the concepts figuring in thought and also which concepts one entertains in thought. For instance, in the case of Rachel in chapter 2, she was not only wrong about what concept she was using, but also the content of her concept tokens. Likewise, switching-Oscar does not know that his concept *water* that originated at Earth has a different content than the concept *water* acquired after the switching. Then it seems that Sainsbury and Tye must make a stronger claim than it first seemed: In addition to abandoning IKCC, which to them is a thesis about meaning vehicles, Sainsbury and Tye must also deny that individuals can know via introspection the representational content of the concepts figuring in thought.

know whether to endorse the thought or to reject it. The puzzle, then, is how one could fail to know whether the statement of identity is true in the thought *that is that*, when *that* refers to the same object in both cases.

Sainsbury and Tye give the following solution to the puzzle: The two occurrences of the concept *that* in the thought *that is that* are of distinct types. Indexical concepts, recall, contrast with other concepts in that individuals can introduce them to themselves independent of other people. Even though both of Jonathan's concepts *that* belong to the same concept template (recall that concept templates are not themselves concepts, but rather rules for forming concepts), the concepts have distinct origins, and thus they are of distinct types, according to originalism. One of the concepts originated when Jonathan intended to refer to what he actually saw with his left eye, while the other originated when Jonathan intended to refer to what he actually saw with his right eye. Since distinct concepts play distinct cognitive roles, this explains how Jonathan can fail to be in a position to know whether the identity statement is true or not.

3.6 The Puzzle of Empty Thoughts

The sixth puzzle, which I have already addressed, is the puzzle of empty thoughts. Concepts such as *Vulcan* and *Pegasus* do not have any reference since, in the actual world, there is no such thing as the planet Vulcan or winged horses. Even so, we can use the concepts in thoughts. For instance, one might think that Vulcan does not exist. Hence, thoughts might involve concepts such as *Vulcan* and *Pegasus,* even if these concepts do not pick out anything in the world. How, then, can such concepts be part of genuine thoughts? Explaining how empty concepts can play an interesting role in cognition is one of the main motivations behind originalism. Since, according to originalism, concepts are not to be individuated in terms of their semantic features, two concepts that are both empty may nonetheless be of distinct types. As we saw in the solution to the puzzle of Hesperus and Phosphorus, referentially isomorphic thoughts may play different roles in cognition. This explains why it is one thing to think that Pegasus is a winged horse and another to think that Vulcan is a winged horse; the concepts deployed in the two thoughts have different origins. Further, even if *concepts* may lack content, according to originalism, every *thought* has a content. A thought is true if and only if the actual world is a member of the associated set of possible worlds. A thought consisting of

only a nominative concept and a predicate, such as *Pegasus is a horse,* is false in all cases where one (or more) of the constituent concepts lack a reference in the actual world. Negative thoughts containing empty concepts, on the other hand, come out as true. For instance, the negative thought *Pegasus is not a horse* is true because it is a negation of the false thought *Pegasus is a horse*, and the negation of a false thought is true. This explains how empty concepts can generate genuine thoughts. Here again, the vehicles of content play a similar role to the one modes of presentation does on the Fregean account, without necessitating a further layer of semantics.

3.7 The Puzzle of Thinking About Oneself

The last puzzle addressed by Sainsbury and Tye is the puzzle of thinking about oneself. Consider the following case: Ernst Mach, not recognizing his own reflection in the mirror as himself, thinks that the man he sees is a shabby pedagogue. However, he does not think that he, himself, is a shabby pedagogue; he does not have the belief *I am a shabby pedagogue* (Mach 1914, 4n.). This implies that there is a difference between thinking of oneself in third person and thinking of oneself in first person. Further, there seems to be two different ways of using the concept *I* (or *my*): I could be mistaken about whether my arm is broken or whether I've grown six inches, but it seems that I cannot be wrong about whether I am in pain. That is to say, it can make sense for someone to ask if I am sure whether it is really my arm, and not someone else's, that is broken, but to ask whether I am sure that it's me, and not someone else, that experiences the pain seems absurd.[42] Does the latter way of thinking about the subject engender some special immunity to error? According to originalism, Mach has more than one concept that refers to himself: one of them being *I* another one being *that shabby pedagogue.*[43] Both concepts have the same content, according to originalism, since they are co-referential. Mach's failing to recognize reflection in the mirror as himself is, thus, "essentially like the failure of early astronomers to recognize Hesperus as Phosphorus" (Sainsbury and Tye 2012, 144). This explains why Mach can be rational in thinking that the person he sees in the mirror is a shabby pedagogue while doubting that he himself is a shabby pedagogue; the relevant thoughts contain distinct concepts. Further, the I-template is governed by the rule that the tokened concepts are to be used to think of oneself. Acquiring a

[42] This point is originally due to Wittgenstein (1958).

[43] Note that the concept *I* is atomic, whereas *that shabby pedagogue* is non-atomic.

certain I-concept involves mastering this rule. Your concept *I* can never refer to anybody but you; hence self-reference is guaranteed to succeed. In contrast, there is no apriori guarantee that *Ernst Mach* refers to the same individual as does Mach's concept *I,* so Ernst Mach may be mistaken about the reference of his concept *Ernst Mach.* This explains why there is an important difference between thinking about oneself in the first person perspective and thinking about oneself in a third person perspective.[44]

3.8 Chapter Summary

Originalism is advanced to solve seven of the classical puzzles in philosophy of language and philosophy of mind. According to Sainsbury and Tye one is in no need of a sophisticated theory of semantics in order to explain these puzzles; instead they make the vehicles of content available as a separate source of explanation. In this chapter I've presented the puzzles with which Sainsbury and Tye are concerned and the originalist solution given to these puzzles. A common feature of the originalist solutions is that they explain the cognitive role of concepts and thoughts in terms of vehicles of content rather than the content expressed by such vehicles. In chapter 4, I will take a critical look at some of these solutions; more specifically, I will discuss the solutions given to the puzzle of Paderewski, the puzzle of Hesperus and Phosphorus and the puzzle of twins, and argue that Sainsbury and Tye fail to solve these puzzles. I will show that the originalist solution to Kripke's Paderewski puzzle gives rise to serious problems and that it does not succeed in accounting for Peter being rational in holding seemingly contradictory beliefs. Further, I will argue that Sainsbury and Tye fail to explain what was central to Frege's original puzzle: the issue of how one can make empirical discoveries by coming to know informative identity statements. I also argue that the originalist solution to the puzzle of twins does not really add anything to the classic debate about mental content. If I am right, and Sainsbury and Tye's originalist account in fact fails to solve the specific problems it is advanced to solve, this casts doubt on the theory as a whole.

[44] Sainsbury and Tye give a more extensive discussion of this puzzle than I present here. I choose not to go into too much detail about this puzzle, since nothing of what I say later on bears directly on the originalist account of thinking about oneself.

Chapter 4

Puzzles left Unsolved

Sainsbury and Tye suggest originalist solutions to seven classical puzzles in philosophy of mind. If their suggested solutions do not work, we shall have reasons to reject the view. In this chapter, I will focus on three specific puzzles and demonstrate that originalism is not suited to solve them. First, I criticise the originalist solution to the puzzle of Paderewski and show that Peter's being rational cannot be explained by appealing to his having false second order beliefs. I then take a closer look at Sainsbury and Tye's solution to the puzzle of Hesperus and Phosphorus and argue that they fail to explain Frege's initial puzzle, since they cannot explain how one can make an empirical discovery through informative identity statements. Finally, I argue that the originalist account of the puzzle of twins does not really contribute to the classic debate going back to Putnam (1975) and Burge (1979b). Again, if my arguments are successful, they pose serious problems for originalism since solving the relevant puzzles is the *raison d'être* of originalism.

4.1 The Puzzle of Paderewski Revisited

The first puzzle I will take a closer look at is Kripke's puzzle about Paderewski. The case, recall, involves Peter having both the belief *Paderewski has musical talent* and the belief *Paderewski does not have musical talent* about a single person Paderewski, while taking there to be two such individuals: Paderewski (the musician) and Paderewski (the politician). In response to this case, Sainsbury and Tye claim that Peter has a false second order belief; he falsely believes that he does not believe that Paderewski has musical talent when forming the

belief that Paderewski lacks musical talent at the political rally: "Peter is asked if Paderewski has musical talent. In responding negatively, he is being sincere; sincerity means he believes what he says; what he says is that he does not believe that Paderewski has musical talent; so we can infer that he believes that he does not believe that Paderewski has musical talent" (Sainsbury and Tye 2012, 137). This in turn explains how he can be rational in holding contradictory beliefs. I will show that this solution is untenable; it leads to Peter having more sets of contradictory beliefs than in the original puzzle. Consider the following: One day Peter is asked the same question – whether Paderewski has musical talent – by one of his music loving friends that Peter thinks knows nothing about politics. This time Peter gives a positive response. The case just portrayed can be described in the same way as the case addressed by Sainsbury and Tye above: When giving the positive response Peter is being sincere; he believes what he says; what he says is that he believes that Paderewski has musical talent; so we can infer that Peter believes that he believes that Paderewski has musical talent. But we have already inferred that Peter believes that he does not believe that Paderewski has musical talent. It seems that Peter's second order beliefs are contradictory: Peter believes that he does not believe that Paderewski has musical talent and he also believes that he believes that Paderewski has musical talent.[45] The appeal to higher order beliefs, then, results in us having to attribute to Peter contradictory second order beliefs and thus this route fails at explaining Peter's being rational.

Let me expand this. It is part of the story that before making the discovery that Paderewski (the musician) = Paderewski (the politician), Peter believes both that Paderewski has musical talent and also that Paderewski does not have musical talent. We need to explain how Peter can be rational in continuing believing that Paderewski has musical talent after having formed the belief that Paderewski lacks musical talent. Following Sainsbury and Tye we can appeal to Peter having a false second order belief. But consider the following. When attending a new concert with Paderewski, Peter is asked by his friend whether he believes that Paderewski lacks musical talent. This time Peter gives a negative answer. In responding negatively, he is being sincere; he believes that he does not believe that Paderewski lacks musical talent.

[45] We may assume that Peter's second order beliefs are occcurrent beliefs. If someone was to object that this seems implausible due to the long timespan between Peter's forming the first and the second belief, we could just adjust the story: Instead of Peter being asked whether he thinks Paderewski has musical talent at different events, two of Peter's friends – one of which is a music lover and who knows nothing about politics, and the other a politician who knows nothing about music – could ask him whether he believes that Paderewski has musical talent at the same time, and Peter would still give one positive and one negative answer. We could thus infer that Peter has both second order beliefs at the same time.

Hence, we can infer that Peter has a false second order belief that accounts for him being rational in continuing believing that Paderewski has musical talent. Now, an hour later Peter is talking to a friend of his that is a politician and that Peter thinks knows nothing about music. During their conversation Peter is asked the same question; whether he believes that Paderewski lacks musical talent. This time Peter responds positively; he believes that he believes that Paderewski lacks musical talent. But, then, again Peter ends up having contradictory second order beliefs: Peter believes that he does not believe that Paderewski lacks musical talent and he also believes that he believes that Paderewski lacks musical talent. It seems that we must ascribe to Peter an additional set of contradictory second order beliefs in order to explain why he is rational in continuing believing that Paderewski lacks musical talent after having formed the belief that Paderewski *has* musical talent. But now the situation is this: Peter believes (i) that he has the belief that Paderewski has musical talent; (ii) that he does not have the belief that Paderewski has musical talent; (iii) that he has the belief that Paderewski lacks musical talent, and (iv) that he does not have the belief that Paderewski lacks musical talent. Belief (i) and (ii) are contradictory; one asserts what the other denies, and the same is the case with (iii) and (iv). We end up with a new puzzle: How can it be that Peter, being rational, can have two sets of seemingly contradictory second order beliefs?

Following the solution given to the Paderewski case in *Seven Puzzles of Thought,* we could try appealing to Peter having false third order beliefs about his second order beliefs. But this is no good, since this route just renders the case to be such that Peter has even more contradictory beliefs; four sets of contradictory third order beliefs, to be exact – two sets for each of the contradictory second order beliefs. This would then turn into an endless regress, producing yet more contradictions further up the chain. Hence, trying to explain why Peter is rational in having contradictory beliefs at one level by appeal to false higher order beliefs, renders us with more sets of contradictory beliefs the further up we go. Hence, Sainsbury and Tye's solution to the Paderewski case fails to explain why Peter is rational in believing (P & ¬P), since their solution creates more problems that threaten Peter's being rational. How can originalists respond?

Appealing to a difference in type of the concept tokens used by Peter is not an option for Sainsbury and Tye, since they hold that concepts are public and it is only one public concept *Paderewski.* Even if Peter thinks there are two individuals called 'Paderewski', Sainsbury and Tye must say that he only uses one concept *Paderewski*, since they are independently

committed to saying that Peter being wrong about the reference of *Paderewski* does not prevent him from using the public concept: "Concept possession is consistent with all sorts of mistakes and misunderstandings about the concept's subject matter" (Sainsbury and Tye 2012, 55). Hence, as noted, originalism must appeal to something else than their core assumptions in order to explain the puzzle of Paderewski. As I have shown, appealing to Peter having false second order beliefs is misguided, since this route renders it even harder than in the original case to explain Peter's being rational. Hence, Sainsbury and Tye's solution to the puzzle of Paderewski fails. I will make a related point in 5.3, where I put forth my own thought experiment in order to cast further doubt on originalism as a general theory of concepts and thoughts. I will now go on to discuss their solution to the puzzle of Hesperus and Phosphorus and argue that Sainsbury and Tye also fail to explain that puzzle.

4.2 The Puzzle of Hesperus and Phosphorus Revisited

Before I turn to the main criticism of Sainsbury and Tye's solution to the puzzle of Hesperus and Phosphorus, I will make a few comments about their solution. As noted, Sainsbury and Tye hold that the difference in informativeness in identity statements is to be explained by a difference in cognitive processing due to the number of concepts figuring in thought. According to originalists, the number of concepts deployed in thought depends on the origins of the relevant concepts and not on the content expressed by such entities. Even if two concepts share content, the concepts are processed as distinct concepts if they have distinct origins. This supposedly explains why it is one thing to think that Hesperus is Hesperus and another to think that Hesperus is Phosphorus. In the first case, the same concept is active twice, so that by the time of the second tokening the process effort needed has already been made. In the second case, however, there are two distinct concepts so more effort is needed in order for the brain to process the thought. Now, Sainsbury and Tye's solution to the classic Frege puzzle presupposes that *Hesperus* and *Phosphorus* have distinct origins; if the concepts had the same origin, they would be of the same type, according to originalism, and then it seems that the cognitive processing needed for the thought *Hesperus is Phosphorus* would be the same as in the case of the trivial thought *Hesperus is Hesperus*. However, there is no apriori guarantee that *Hesperus* and *Phosphorus* have distinct origins. Even though it seems historically unlikely, it may turn out that *Hesperus* and *Phosphorus* have the same origin and

that their difference in spelling and pronunciation developed gradually through time[46]. If this is the case – if *Hesperus* and *Phosphorus* turn out to be the same concept – the originalist solution to the puzzle is ineffective; there would be no real difference in the cognitive significance of *Hesperus is Hesperus* and *Hesperus is Phosphorus.* For the originalist solution to be effective, then, it seems that one must stipulate that *Hesperus* and *Phosphorus* have distinct origins, but this might turn out to be historically incorrect.

If it should turn out that *Hesperus* and *Phosphorus* do in fact have the same origin, the thought *Hesperus is visible* would be the same as the thought *Phosphorus is visible,* both at the level of content and vehicles. However, part of the Fregean data Sainsbury and Tye set out to explain is exactly that it is one thing to think that Hesperus is visible and another to think that Phosphorus is visible. If the origin of *Hesperus* and *Phosphorus* is the same, then, Sainsbury cannot explain this data without appealing to something else than vehicles and referential content. Further, the ancient Babylonians might have formed the beliefs *Hesperus is visible* and *Phosphorus is not visible* at the same time. If *Hesperus* and *Phosphorus* have the same origin, the thoughts would be contradictory even on the originalist framework, since the thoughts contain concepts of the same type and one denies what the other asserts. Still, the Ancient Babylonians would be rational in forming both beliefs since they do not know that the concepts are the same. The situation, then, is similar to the one of Peter in the Paderewski case. If someone is rational in forming thoughts that are contradictory at the level of vehicles, Sainsbury and Tye must appeal to something else in order to explain them being rational; they must appeal to the thinker being wrong about her beliefs due to her being wrong about the nature of the concepts involved in thought. However, if my criticism of Sainsbury and Tye's solution to the Paderewski case is correct, they cannot explain cases in which thoughts that agree in vehicles and content play distinct roles in cognition. Hence, in order to avoid turning the puzzle of Hesperus and Phosphorus into a case along the lines of the Paderewski case, Sainsbury and Tye must stipulate that the concepts have distinct origins. But, as noted, even if it seems historically plausible that *Hesperus* and *Phosphorus* have distinct origins there is no apriori guarantee that they do not have the same origin. I will now turn to my objection to Sainsbury and Tye's solution of the puzzle of Hesperus and Phosphorus and argue that even if they make such stipulations, their solution to Frege's puzzle still fails.

[46] Note that originalists must claim that it is possible that symbols that have different *shapes* may still count as the same concept.

The problem I want to raise for the originalist solution to Frege's puzzle is that it fails to explain how informative identity statements can provide us with knowledge about the world, and not just knowledge about our concepts. Frege's, with his initial puzzle (see 1.1), was concerned with how such identity statements can provide us with empirical knowledge about the world: how they can contain "very valuable extensions of our knowledge" (Frege 1892, 56). According to originalism, the discovery made by the Ancient Babylonians when learning that Hesperus is Phosphorus is a *cognitive discovery*. They learned something about their *concepts*, namely that they expressed the same content – they did not learn that such and such was the case in the external world. The Ancient Babylonians may have formed the thought *Hesperus is Phosphorus* before making the discovery, but they did not believe it, and if they did, they did not have the necessary evidence for the belief being justified. What happened when they came to know that Hesperus is Phosphorus, according to Sainsbury and Tye, is that a new thought came to be knowledge. Since, on this view, the thought *Hesperus is Hesperus* and *Hesperus is Phosphorus* has the same content (that is, they share sets of possible worlds in which they are true), one gets no new information about the world when learning that the latter is true: "In some sense, no new "fact" was added to knowledge when the fact that Hesperus is Phosphorus was added" (Sainsbury and Tye 2012, 125). However, it seems implausible that the ancient Babylonians did not make a discovery about the world, but merely a cognitive discovery, when learning that Hesperus is Phosphorus. This was at the centre of Frege's initial puzzle.

As noted, Sainsbury and Tye model their theory of concepts on words. If it is the case that a theory of concepts really should be understood as corresponding to a theory of words, the following case should be analogous to the case of the Ancient Babylonians discovering that Hesperus is Phosphorus: First, let's say I believe that Tyler is fair. 'Fair' is synonymous with 'impartial'; that is, they have the same meaning (or so I will assume). Now let's say that, for some reason, I have misunderstood the term 'impartial'; I think that 'impartial' means 'unfair'. Because of my misunderstanding I might come to form the belief that Tyler is not impartial. One day I discover that I have misunderstood the terms, and come to form the belief that Tyler is impartial. However, when learning that 'fair' and 'impartial' have the same meaning, I make a discovery about language and not about Tyler. When making the discovery, no new fact about Tyler was added to my knowledge, since I already had the belief that Tyler is fair (and 'fair' and 'impartial' are assumed to have the same meaning). My discovery in this case consists in learning that two words that I thought had different meaning

do in fact have the same meaning. This story is analogous to Sainsbury and Tye's account of the Ancient Babylonians discovering that Hesperus is Phosphorus; the Ancient Babylonians made a discovery about their concepts and not about any new empirical fact. But this is clearly not analogous to Frege's case. To see this, consider the following.

If Lois Lane discovered that Clark Kent is Superman, she would make a discovery about the world, and not just about her concepts. In particular, she would learn a new fact about her colleague Clark Kent.[47] This, unlike the case above, is analogous to Frege's case. Just like Lois Lane makes a discovery about the world when learning that Clark Kent is Superman, so did the Ancient Babylonians when learning that Hesperus is Phosphorus. When learning that Hesperus is Phosphorus they made a discovery about Hesperus (e.g. that it is also visible in the morning). After making the discovery, the Ancient Babylonians were justified in inferring that the same heavenly body is visible twice a day, and this is an empirical discovery. How can we explain someone learning a new fact about the world from knowledge only about meaning vehicles? Frege made this point in the opening paragraph of his (1892), where the puzzle is raised. In his (1879), Frege suggested that informative identity statements result only in cognitive discoveries, but he abandons this idea in his (1892) when confronted with the puzzle. Again, the puzzle concerns how one can gain knowledge about the world from such statements. Obviously, responding that only a cognitive discovery is made is no good: it begs the question, and offers no explanation of how empirical knowledge can be gained. A solution to Frege's puzzle cannot just be an account of how to distinguish trivial identity statement from informative identity statements, but must also provide an explanation as to why thoughts such as *Hesperus is Phosphorus* are informative in the sense that from them we get new knowledge about the world. Sainsbury and Tye provide no such explanation and hence the theory fails at solving Frege's puzzle.

4.3 The Puzzle of Twins Revisited

The final case I want to take a closer look at is Sainsbury and Tye's solution to the Twin Earth case of Putnam (1975). Sainsbury and Tye explain the difference in the thoughts of Oscar and Twin Oscar by appealing to a difference in concepts; since Oscar's concept *water*

[47] Joseph Hedger (forthcoming) makes the same point regarding originalism failing to explain Frege's puzzle. I borrow the illustration of the comparison between true synonyms and co-referential concepts from him.

is distinct from Twin Oscar's concept *water,* their thoughts as a whole are distinct. The thought *that is water*$_1$ entertained by Oscar plays a different cognitive role than Twin Oscar's thought *that is water*$_2$*,* since the thoughts contain the distinct concepts *water*$_1$ and *water*$_2$ (one originating at Earth, the other at Twin Earth) and since cognitive significance is to be explained in terms of meaning vehicles.. When laying out the puzzle of twins, Sainsbury and Tye present the classic puzzles of Putnam and Burge regarding externalism and internalism about mental content, implying that originalism can make a contribution to this debate. However, while originalism explains how the twins' thought vehicles can be distinct, it does not really add anything to the standard externalist theories of mental content. This is reflected in the key assumptions of originalism being compatible with both externalism and internalism about mental content. Historically the relevant debate about internalism and externalism is about how intrinsic duplicates can differ with respect to their *mental content*, and not how they can differ with respect to their meaning vehicles. This is what the traditional puzzles are concerned with. Let me show the irrelevance of the originalist solution to these puzzles.[48]

Traditional externalists hold that Oscar and Twin Oscar have distinct mental contents when thinking about water. However, most traditional externalists would want to say that when thinking of something that has the exact same intrinsic features on the two planets, Oscar and Twin Oscar share mental content. That is to say, the difference in mental content of Oscar and Twin Oscar is due to them standing in relations to liquids with distinct chemical structure; had the watery stuff on Twin Earth turned out to be made up of H_2O, just like water on Earth, Oscar and Twin Oscar would have had the same mental content (and, as a result, be in the same mental state). According to originalism, however, in this case Oscar and Twin Oscar still have distinct *water* concepts, since one originated at Earth and the other at Twin Earth. Their thoughts still play different roles in cognition since they contain distinct concepts. Sainsbury and Tye do not state explicitly that they think inhabitants on Earth and Twin Earth can share mental content, but there are reasons to believe that they would agree with this. According to Sainsbury and Tye, the content of concepts is, in most cases, introduced at the origins of a concept and then maintained through time because of the mechanism of deference. In most cases, then, the content of concepts depend on the intentions of the person who introduced the concept.

[48] It might seem here that I am begging the question against Sainsbury and Tye. They might claim that they are setting out to solve a different puzzle, one about how the twins can differ at the level of vehicles. But it is clear from their presentation that they set out to solve the traditional puzzle.

For instance, when Gell-Mann introduced the concept *quark* he intended the concept to pick out quarks in the entire universe, and not just on Earth. Since, in Putnam's thought experiment, Twin Earth is stipulated to be a distant planet in our own universe, Gell-Mann's concept also picks out quarks on Twin Earth. The same is the case for Gell-Mann's intrinsic duplicate Twin Gell-Mann; Twin Gell-Mann's concept *quark* also refers to all quarks in the universe. Hence, when Gell-Mann and Twin Gell-Mann use their distinct concepts *quark* they have the same representational content: When Gell-Mann has a belief containing his concept *quark* and Twin Gell-Mann has an equivalent belief only with his concept *quark*, their beliefs have the same truth conditions. Hence, it seems plausible that Sainsbury and Tye would agree that someone on Earth and someone on Twin Earth could share mental content. However, their theory does nothing to accommodate the difference between intrinsic duplicates having the same or different mental content: Every concept on Earth is distinct from equivalent concepts originating on Twin Earth. It seems plausible that Gell-Mann's thought *quark are tiny* plays the same cognitive role as Twin Gell-Mann's thought *quarks are tiny*, but their concepts are distinct, according to originalism. Since the concepts figuring in the two thoughts are distinct the thoughts should play distinct roles in cognition. In this case, however, it seems as though their thoughts *do* play the same role in cognition, but this cannot be explained by appeal to their meaning vehicles individuated in terms of their origins. Rather, in this case, it seems like their sameness of representational content is what explains them having thoughts with the same cognitive significance. Then it seems that what accounts for Oscar and Twin Oscar being in different mental states, whereas Gell-Mann and Twin Gell-Mann are in the same mental state is the content of their thoughts, and not the vehicles. In this case, a difference in concepts does not suffice to explain what is special about the relation between the concepts $water_1$ and $water_2$ as opposed to *quark* and *twin quark.* Hence, the solution suggested by Sainsbury and Tye does not contribute to the classic puzzle of twins, since there is still the question of how intrinsic duplicates can differ with respect to their mental content.

The originalists might want to argue that distinct concepts *can* play different roles in cognition, but that they don't have to. Could they perhaps say that concepts may play the same role in cognition if they have qualitatively identical origins? For instance, they could claim that *quark* and *twin quark* play the same cognitive role due to the concepts having qualitatively identical origins. *Water* and *twin water*, on the other hand, do not have qualitatively identical origins since the introduction of each concepts involved the people

using them for the first time standing in a relation to qualitatively distinct entities (H_2O and XYZ respectively). This response, however, is not available to originalists. The reason is that, according to the view just outlined, features other than the mere point in history at which the concept was first introduced and the intention of the person using the concept intentionally for the first time, are essential to the role a concept plays in cognition. On this route, the difference in cognitive significance of *water*$_1$ and *water*$_2$ is to be explained in terms of the relation the individual introducing the concept has to her environment. But originalists deny that such features are needed to explain the cognitive role of concepts and thoughts. Hence, the route just outlined is not available to the originalist. In chapter 7 I present a theory that in many respects resembles originalism, but according to which the relation individuals bear to objects in their environment is essential for the individuation of their concepts. I will argue that this alternative view, Recanati's theory of mental files, is preferable to originalism in that it avoids the problems posed for originalism in this thesis and at the same time preserves appealing aspects of originalism. Let me now return to the traditional twin cases, however.

It is striking that in their presentation of twin puzzles Sainsbury and Tye put forth Tyler Burge's (1979b) thought experiment in favour of externalism about mental content, but when giving a solution to the puzzle, they only address the Twin Earth puzzle, stemming from Putnam. In his thought experiment, Burge highlights the relevance of the thinker's relations to her social community for the individuation of mental content. Sainsbury and Tye agree with Burge that one's language community is important for the content of concepts and thoughts; the content of such entities is determined by the public use of the group to which one defers.[49] Burge puts forth a thought experiment that is supposed to establish that a person can maintain his intrinsic properties, but at the same time have different mental content as a result of alterations in his environment: Alf has several beliefs about the illness arthritis, many of which are true. He does, however, believe that he has arthritis in his thigh, which is false since arthritis may only affect the joints. Now, let's imagine a counterfactual world in which 'arthritis' is a more general term, which also includes the disease in Alf's thigh. In this world, Alf's belief that he has arthritis in his thigh is true. Burge claims that in the counterfactual world Alf has no concept of arthritis, and thus he has no beliefs about arthritis. Instead he has the concept *tarthritis.* Alf's physical and non-intentional state is identical in the counterfactual world and the actual world, but Alf has different intentional content in the actual world and

[49] Sainsbury and Tye do, however, disagree with Burge's account in that they find no use for a notion of *experts* when explaining the public meaning of concepts.

the counterfactual case; *arthritis* and *tarthritis* are different concepts (on the traditional understanding of 'concepts', according to which such entities are constitutively semantic), according to Burge. Thus, Burge concludes that a thinker's intentional state depends on certain features of her linguistic environment.

The puzzle Sainsbury and Tye draws from Burge's thought experiment is the same as the one drawn from the Twin Earth case: How can intrinsic duplicates differ with respect to their thoughts (i.e. vehicles of content)? Since Sainsbury and Tye do not address this case directly when solving the puzzles, I can only make a guess as to what their solution would be. From their formulation of the puzzle, it seems to be the case that Sainsbury and Tye think Alf in the actual case has a different thought than Alf in the counterfactual case when thinking *I have arthritis*. According to originalism, in order for the thoughts to be distinct, they must contain concepts with different origins since they have the same structure. Hence, *arthritis* and *tarthritis* must have distinct origins, according to originalism. However, Burge's thought experiment is silent on how *arthritis* and *tarthritis* first came into being. Is it necessarily the case that the two concepts have distinct origins? One could, of course, stipulate that in the counterfactual case, *tarthritis* has a different origin than *arthritis* has in the actual case, since they have different content. The person first introducing the concepts may have introduced them at distinct points in history in the actual and counterfactual case; it might not even have been the same person who first introduced the concepts. In this case, the concepts are distinct, according to originalism. However, another possibility is that *arthritis* and *tarthritis* have the same origin in the actual and counterfactual case. Since originalism allows for concepts to change their content, it may be the case that the difference in content between *arthritis* and *tarthritis* developed through time. Initially both concepts may have included only inflammation of the joints, but then gradually, *tarthritis* came to pick a wider range of illnesses. In this case, *arthritis* and *tarthritis* are the same concept, according to originalism, since they have the same origin. If the concepts are the same, the thought *I have arthritis* should be the same for Alf in the actual case and Alf in the counterfactual case. Since Burge is silent on how the origins of *arthritis* and *tarthritis* and also how they came to have their specific content, we have no reason to prefer one of the stories presented to the other. In order for the originalists to hold that Alf has distinct thoughts in the actual and counterfactual case, then, they must assume that *arthritis* and *tarthritis* have distinct origins. But then, once again, Sainsbury and Tye explanation rest on assumptions that are not necessarily true. This casts

further doubts about Sainsbury and Tye's account as a general theory of the nature of concepts and thoughts.

One possibility for Sainsbury and Tye might be to allow for Alf having the same thought in the actual and counterfactual case. They could say that Alf's thoughts differ with respect to content but not at the level of meaning vehicles. This, however, seems like an unmotivated response. In the Twin Earth case, Sainsbury and Tye are clear that Oscar and Twin Oscar have distinct concepts – in fact they must say this, since the concepts have distinct origins. This is what their contribution to the debate about internalism and externalism in philosophy of mind amounts to. For someone endorsing social externalism about mental content, such as Sainsbury and Tye, it seems unmotivated to say that Oscar and Twin Oscar have distinct thoughts while Alf has the same thought in the actual and counterfactual case. Indeed, the puzzle they draw from the traditional cases is: *How can it be that the subjects differ with respect to their thoughts?* Needless to say, claiming without further explanation that they do not does nothing to solve the puzzle. Also, if holding that Alf has the same thought in the actual and counterfactual situation, the originalist account provides even less of a solution to the classical puzzle; we seem to be at the same place we were pre originalism. We still need to explain how intrinsic duplicates can differ with respect to their mental content as a result of a difference in their linguistic community. Based on the way the puzzle is set up, however, I suspect originalists would prefer the first route, saying that *arthritis* and *tarthritis* are distinct concepts. The first route is better in the sense that it aims at a contribution to the classical debate, but the route rests on stipulations that are not apriori guaranteed. Further, even if such stipulations are made, the originalist solution does not solve the puzzle of how intrinsic duplicates can differ with respect to their mental content, which is what the original puzzle of twins is really about.

4.4 Chapter Summary

The reason given to believe originalism is that the theory supposedly solves the classical puzzles of thought: "Which view is ultimately superior depends on which view offers the best account of the various puzzles of thought" (Sainsbury and Tye 2012, 115). Hence, the plausibility of originalism turns on its ability to solve the classical puzzles. In this chapter I have argued that the originalist solutions to three of the puzzles are unsatisfactory. I started

out by arguing that the solution given to the puzzle of Paderewski fails. The reason is that an appeal to false higher order beliefs generates even more sets of contradictory beliefs, threatening the account of Peter's being rational. One cannot explain someone being rational in holding contradictory first order beliefs if one then ends up having to attribute even more (in fact an infinite number of) sets of contradictory higher order beliefs to the individual.

I then argued that Sainsbury and Tye's solution to the puzzle of Hesperus and Phosphorus fails. This is because they cannot explain how identity statements can be informative in the sense that they potentially provide new information about the world. They are committed to the view that such discoveries only amount to cognitive discoveries. But explaining how informative identity statements potentially involve making a discovery *about the world* is at the core of the original puzzle stemming from Frege. Sainsbury and Tye thus fail to explain the central feature of the puzzle of Hesperus and Phosphorus.

I also argued that the originalist interpretation of the puzzle of twins is misguided. I then argued that their solution to the original puzzle is irrelevant, and that their solution to the puzzle they draw from the original puzzle depends on serious stipulations and fails to reflect the difference between intrinsic duplicates having the same mental states on the one hand and their having distinct mental states on the other. Sainsbury and Tye take the puzzle of twins to be how intrinsic duplicates can differ with respect to their thoughts (vehicles of content). The classic debate, however, is concerned with how intrinsic duplicates can differ with respect to their mental *content*. Also, while originalism might be able to explain the difference in thoughts of inhabitants of Earth and Twin Earth, they fail to explain the sameness in thought in cases where the thoughts have the same content. More precisely; for originalists it is of no importance that the watery stuff on Twin Earth is made up of XYZ; Oscar and Twin Oscar would have had thoughts of different types even if the watery stuff on Earth and Twin Earth had the exact same molecular structure. Hence, the originalist solution to the classic puzzle of twins is extremely limited, and the limitation is reflected in the key claims of originalism being compatible with both internalism and externalism about mental content. Furthermore, I showed that originalism cannot claim that Alf (from Burge's (1979b) case) in the actual and counterfactual case have distinct thoughts without stipulating that *arthritis* and *tarthritis* have distinct origins. However, since the original thought experiment is silent on how the concepts originated, we have no reason to prefer this story to one in which *arthritis* and *tarthritis* have the same origin. If they have the same origins, Alf's thoughts would be the same in both

cases, but the puzzle they pose based on Burge's case is precisely how it is that the thoughts can differ. The originalist solution thus does not seem to add anything to our theorising.

Solving the seven puzzles of thought is the *raison d'être* of originalism. Hence, if I am right that originalism fails to solve the puzzles addressed in this chapter, this casts doubt on the theory. In the next chapter I will challenge originalism further. I will put forth a thought experiment that is supposed to establish that originalism about concepts and thoughts cannot give a sufficient explanation of rationality – which is one of the central explanatory roles of thoughts. The upshot of the thought experiment is that, if taking concepts to be public and individuated by way of their origin, this cannot provide a general account of the cognitive role of thoughts. Since in *Seven Puzzles of Thought* Sainsbury and Tye are mostly concerned with the seven puzzles, and give no explicit statement of what a general theory of concepts and thoughts would be, I make certain suggestions to this on their behalf. More precisely, I propose two different approaches and argue that both fail to explain certain cases of rationality.

Chapter 5

Originalism and Rationality

One of the central explanatory roles of thoughts is to explain rational cognition. Any general theory of thought must be able to accommodate this. In this chapter we shall see how Sainsbury and Tye's originalism deals with this task; it appears that their theory cannot explain certain cases of rationality, or so I will argue. First, let us bring to mind Sainsbury and Tye's central claims:

> Atomic concepts are to be individuated by their historic origins, as opposed to their semantic or epistemic properties. Distinct concepts have different origins, and may not differ intrinsically. Originalists reject the view that cognitive differences need to be explained by semantic differences. […] Individuating concepts in a non-semantic way shows that the explanation does not rely on semantic properties of the concepts themselves (Sainsbury and Tye 2012, 40).

As they present their account in *Seven Puzzles of Thought,* it is not clear exactly what view Sainsbury and Tye take on cognition in general. Their explanation of cognitive significance in terms of cognitive processing effort together with the claim that thinkers only stand in relation to syntactic features and not directly to the semantic properties of thoughts, indicates that their account resembles that of Fodor in his (1980), where he advocates his version of CTM. There are, however, some important differences between the two theories. For instance, Fodor argues that proponents of CTM are committed to saying that mental states can only be distinct if the syntactic features are distinct: "Fix the subject and the relation, and then mental states can be (type) distinct only if the representations which constitutes their objects are formally distinct" (Fodor 1980, 652). That is, there cannot be equivocality in the language of thought. Applied to the originalist framework, this is the claim that a difference in reference indicates a difference in vehicles. Sainsbury and Tye, on the other hand, reject this view when they say

that concept tokens of the same type may differ semantically. However, if what accounts for cognitive significance are the vehicles of content, it seems plausible that two mental states that differ in content must also differ at the level of vehicles. I will show that if the correct understanding of their theory is that cognitive significance essentially depends on the vehicles rather than content, they fail to account for certain cases of cognitive significance if allowing a difference in reference without a difference in vehicles in the mind of a single individual. If thinking doesn't involve standing in a relation to the content of thoughts, we should expect that if there are differences between two mental states due to a difference in representational content, this difference should be reflected in what thinking *does* involve standing in a relation to; the vehicles of content. In the first part of this chapter I will put forth a thought experiment that shows that a result of Sainsbury and Tye' account is that it is possible for concept tokens to express distinct (even contradictory) content within one individual and still be of the same type at the level of vehicles, and that this causes severe difficulties for the explanation of rational cognition in these cases.[50]

It might be objected that Sainsbury and Tye are not committed to the strong claim that every instance of cognitive significance must be explained by appeal to vehicles alone. Even though they explain the seven puzzles in terms of vehicles, it might be that when giving a general theory of cognitive significance they can appeal directly to the reference of thoughts in order to explain cognitive significance. If this is the case, they can avoid the criticism I put forth in 5.2. However, in 5.3 I expand the thought experiment in order to argue that even if Sainsbury and Tye can appeal to semantics when accounting for cognitive significance, there are still cases that cannot be solved within their framework. They fail at explaining rational cognition. First, I present the thought experiment.

5.1 The Thought Experiment

Consider the following case. At some point in history (1300 AD, say) an individual uses the concept *ohun* for the first time. This individual intends to use the concept to pick out anything

[50] The thought experiment is developed as a criticism of originalism in particular and is in no way intended as an argument against theories holding that cognitive significance can be explained in terms of syntactic features in general. Giving a general argument against all such theories is beyond the scope of this thesis; here I will simply explore and criticise the consequences of Sainsbury and Tye's formulation of originalism.

that is an object (much like our concept *object*); for her, if something is an object it is an ohun. As the years go by, this use of the concept *ohun* becomes standardized in the individual's community. Anybody using the concept *ohun* intends to use it the same way as the others in their language community. When someone in the fifteenth century uses the concept *ohun,* she is deferring to earlier uses of the concept and the chain of deference goes all the way back to the point in history at which the concept was intentionally used for the first time. According to originalism, then, everybody in this language community is using the same concept *ohun* as did the first person to intentionally introduce the concept.

In 1500 two groups of people decide to emigrate from the *ohun*-using community. One of the groups, call it group A, settles down on the west coast, while the other group, call it group B, settles down at the east coast. For five hundred years the two groups have no interaction with each other or their old community. As it happens, the two groups' concept *ohun* gradually change their reference. Group A's concept gradually comes to pick out only blue objects. The shift of reference happens without any intentional refixing of the reference or any mayor deviations from standard use being made regarding the concept's reference. This is analogous to the case addressed by Sainsbury and Tye when they account for the change of reference of the concept *meat* (see 2.4). *Meat* originally picked out anything edible, but as time went by, the concept gradually changed its reference as to pick out only animal flesh. According to the originalists, this is not an instance of conceptual fission, since the development happened gradually and no use qualifying as an originating use. Instead it is an instance of a concept staying the same but with a change in reference. When people today use the concept *meat* to pick out animal flesh they use the same concept as did the 1500th century people using the concept to pick out anything edible: "our concept *meat* is the same as the earlier concept *meat.* The basis for this is simply the smooth history, the concept being handed on in ways that unquestionably make new users count as users of the concept, even if mistakes were made about its referent" (Sainsbury & Tye 2012, 72). In my thought experiment the change in reference of the concept *ohun* happens in the exact same way: There is a smooth history and at no point did anybody not count as users of the concept. As in the case of *meat,* small deviations from standard use of the concept *ohun* occur along the way, but none of the mistakes are noticeable and sufficient for introduction of a new concept (see 2.4 for originalism and change in reference). Hence, Sainsbury and Tye must say that group A uses the same concept *ohun* as did the community they emigrated from 500 years earlier; the chain of deference goes all the way back to the point in history at which *ohun* was first used to pick

out anything edible.

As it happens, group B also has a change in the reference of the concept *ohun*: Gradually *ohun* comes to pick out any object that is *not* blue. This shift in reference is completely analogous the change in reference of *ohun* in group A: The shift happens without any intentional refixing of the reference or any mayor error being made about the concept's reference. Hence, following Sainsbury and Tye's handling of the concept *meat*, group B is using the same concept *ohun* as did the community they emigrated from. When someone in group B is using the concept *ohun*, the chain of deference goes all the way back to the point in history at which *ohun* was first used to pick out any object.

Now, the originalists must say that group A and group B are using the same concept *ohun* since their concepts have the same origin; the point in history at which *ohun* was first used to pick out anything that is an object: "Concept C1 = concept C2 iff the originating use of C1 = the originating use of C2" (Sainsbury &Tye 2011, 4). One of the key claims of originalism, recall, is that the origin of a concept is the origin of one concept only, so there can only be one concept *ohun* that is used both by group A and group B. But remember that the *reference* of the concept is distinct when tokened by individuals in the two groups:[51]

Group A: *ohun* refers only to objects that are blue
Group B: *ohun* refers only to object that are not blue

For simplicity, I will refer to group A and group B's use of the concept *ohun* as $ohun_{(A)}$ and $ohun_{(B)}$ respectively. It is important to keep in mind that the difference in notation only reflects a difference in reference, and not a difference in concepts. At no point do the individuals in group A or group B use the notation themselves; they simply use *ohun*.

Now, even though $ohun_{(A)}$ and $ohun_{(B)}$ are the same concept, according to originalism, they express incompatible content. A rational individual can never knowingly use the concept *ohun* to pick out both only blue objects and only objects that aren't blue by the same concept token: If something is correctly taken to be $ohun_{(A)}$ it cannot also be $ohun_{(B)}$; something cannot both be a blue object and an object that isn't blue at the same time. Hence, the

[51] According to Frege, the mode of presentation determines the reference of concepts and thoughts, so a difference in reference implies a difference in mode of presentation. In contrast, originalists take concepts to be individuated by their origins alone, so according to originalism, a difference in reference does not necessitate a difference in concepts and thought.

thoughts *that is an ohun$_{(A)}$* and *that is an ohun$_{(B)}$*, where the indexical expression 'that' refers to the same object, can never both be true at the same time; the sets of worlds in which the two thoughts are true are not overlapping.

I will now take the thought experiment one step further. One day one of the members of group A goes on a journey. After a while this individual, let's call her Indira, settles down on the east coast together with group B. Gradually she becomes aware of group B's different use of the concept *ohun.* Importantly, according to originalism she does not acquire a new concept upon her arrival: Group A and group B share the concept *ohun* since their concepts have the same origin. Instead she acquires a new way of using her old concept. In addition to using *ohun* to pick out only blue objects, Indira now comes to use the concept to pick out only objects that are not blue. When interacting with people in group B Indira defers to the standard use of *ohun* in this language community, and hence use the concept to pick out only objects that are not blue. When talking on the phone or writing letters to people from group A she defers to their use of the concept *ohun* and therefore use the concept to pick out only blue objects. When she is not interacting with any other persons, but forming thoughts in her own company, she can use either *ohun*$_{(A)}$ or *ohun*$_{(B)}$, depending on whom she defers to. Further, Indira being an originalist herself and an enthusiastic researcher of the history of concepts, discovers the common origin of *ohun*$_{(A)}$ and *ohun*$_{(B)}$. Hence, she is well aware that the two groups share the same concepts *ohun,* but with different references. Indira is in a position to think such thoughts as *that is an ohun$_{(A)}$, that is an ohun$_{(B)}$, that is not an ohun$_{(A)}$* and *that is not an ohun$_{(B)}$*. Of course, if *that* refers to the same object, all of these thoughts cannot be true at the same time, but that does not prevent Indira from having the disposition to form such thoughts since thoughts can be false. In 5.2 and 5.3 I will show that the thought experiment just outlined gives rise to worries regarding the explanation of rationality. The upshot is that a theory that takes cognitive significance to rely directly on vehicles of content individuated by way of their origins, cannot explain certain cases of someone being rational or irrational. First, in 5.2, I will show that Sainsbury and Tye cannot give a general account of rationality without taking cognitive significance to depend directly on semantic features of concepts and thoughts if at the same time allowing that concepts of the same type may come to express contradictory content.

5.2 Equivocal Concepts and Rationality

After settling down with group B, Indira's use of the concept *ohun* becomes equivocal:[52] She may use the concept either to refer to blue objects or objects that are not blue, depending on the context (i.e. to whom she defers). When asked by someone in group B if she believes that sapphires are *ohuns*, Indira gives a positive response. When asked the same question during a phone call with someone from group A, Indira gives a negative answer. It then seems corrects that Indira has the belief *sapphires are ohuns*$_{(A)}$ and also that she has the belief *sapphires are not ohuns*$_{(B)}$. Both of these thoughts are true, since sapphires are in fact blue. The set of worlds in which the thoughts are true is the same: In every world where the thought *sapphires are ohuns*$_{(A)}$ is true, necessarily the thought *sapphires are not ohuns*$_{(B)}$ is also true, and vice versa. Hence, intuitively it seems exceedingly plausible that Indira is able to entertain both beliefs at the same time and still be rational. If we only appeal to the thought (meaning vehicle) and not semantics, we could give the following set of claims regarding Indira's beliefs:

B1. Indira believes that sapphires are ohuns
B2. Indira believes that sapphires are not ohuns

Both of these claims are true, but jointly they seem inconsistent. Recall what Sainsbury and Tye say about the necessary and sufficient conditions for thoughts being contradictory: "inconsistent thoughts are contradictory iff one consists of the other embedded in a concept for negation. If one thought contains a nominal concept, a contradiction must contain the same nominal concept at the corresponding position in the structure" (Sainsbury & Tye 2012, 135). In our case Indira's thought in (B1) is embedded in a concept of negation in (B2), so her beliefs must be contradictory on Sainsbury and Tye's view. Since her thoughts are contradictory (and Indira is well aware of this) it seems that she must be counted irrational, but intuitively she is not.

Further, it seems plausible that Indira cannot have both the belief that *sapphires are ohuns*$_{(A)}$ and *sapphires are ohuns*$_{(B)}$ and still be counted as rational. This is because Indira is well aware that the two thoughts could never both be true at the same time: The sets of worlds in

[52] I use the term 'equivocal' to apply to concepts that are of the same type but that have more than one possible content. In language, being equivocal means that words that are spelled and pronounces the same way may differ in meaning. My use of the term is somewhat stronger; the equivocal concepts I'm concerned with are of the very same type due to having the same origin.

which the two thoughts are true are not overlapping; something cannot be both a blue object and not a blue object at the same time. However, according to originalism the two thought tokens are of the same type since they share the same structure and concepts: All of the concepts figuring in the first thought have the same origin as the corresponding concepts in the other. But then it seems that the two thoughts must play the same cognitive role, according to originalism. But if the two thoughts play the same cognitive role, how can Indira be irrational if having both beliefs? Further, consider the thought *ohuns$_{(A)}$ are ohuns$_{(B)}$*. If originalism is true, the thought involves an identity statement between two concept tokens of the same type.[53] Hence, the statement should be trivial, following Sainsbury and Tye. However, the statement is not trivial, let alone true. It is not trivial that objects that are blue are not blue. Recall how Sainsbury and Tye explain the triviality of identity statements between two concept tokens of the same type: "whatever processing effort is required has already been made by the time the second occurrence of the concept is encountered; the previous interpretive outcome can simply be brought forward" (Sainsbury and Tye 2012, 54). But this seems clearly not to be the case in the scenario just outlined. For Indira to think that something is *ohun*$_{(A)}$ clearly involves other cognitive processes than thinking that something is *ohun*$_{(B)}$, otherwise the thought *ohuns$_{(A)}$ are ohuns$_{(B)}$* would be trivial, and it is clearly not. According to originalism, the only difference between *ohun*$_{(A)}$ and *ohun*$_{(B)}$ is their reference. But then it seems clear that the difference in cognitive processing must be directly linked to the semantic features of the thought. Since Sainsbury and Tye hold that thinkers only are related to content indirectly, via vehicles, it is not clear how they their theory can accommodate this observation.

Further, if Indira wants a blue bike and forms the belief *that bike is an ohun$_{(A)}$* she would be rational in buying that bike. However, had she instead formed the belief *that bike is an ohun$_{(B)}$* she would not be rational in buying the bike if what she wants is a blue bike. But according to originalism, the thoughts are of the same type, since they contain the same concepts structured the same way. Also, since Indira is stipulated to be rational, the two thoughts have different causal powers; the thought *that bike is an ohun$_{(A)}$* would cause Indira to buy the bike, whereas the thought *that bike is an ohun$_{(B)}$* would not cause her to buy the bike, given that she only wants a bike that is blue. The fact that the two thoughts have different causal powers indicates that they must play different roles in cognition. If rationality is explained in terms of vehicles individuated by their origins, however, we cannot explain why the two thoughts have

[53] Note again that the thoughts *are not to be individuated* even partly *in terms of their contents.*

different causal powers in Indira's mind, since, on this view, the thoughts must be of the same type; and hence they should play the same role in cognition. On this picture, what explains Indira being rational – and thus her behaviour – is the semantic properties of the concepts and not their origins.

In the case of *Hesperus is Phosphorus,* recall, the difference in concept types reflects a difference in the thinker's relation to the referent. In the case of *Hesperus is Hesperus* the sameness in concepts reflects a sameness in relation to the reference. In the case of *ohuns$_{(A)}$ are ohuns$_{(B)}$,* then, the concepts being of the same type indicates a sameness of relation to reference. The sameness of relation, however, is directed towards two distinct sets of objects. In certain cases it may make sense to say that someone has the same relation to distinct referents; these are cases in which someone wrongly takes two objects or individuals to be the same (such cases are often labelled inverse Paderewski cases (Recanati 2012, 116)). In the case just outlined, however, Indira is not wrong about the nature of the reference of *ohun*$_{(A)}$ and *ohun*$_{(B)}$, so intuitively Indira does *not* relate to the two references in the same way. Indira's relation to the references, then, is not in proportion with the number of concepts she possesses if one takes concepts to be individuated by their origins. This indicates that Sainsbury and Tye cannot give a general account of cognitive significance in terms of vehicles alone, since the way a thinker relates to the content of her thought may not be reflected in the number of concepts deployed in thought.

Let me illustrate further. Deductive reasoning involves reaching a conclusion by way of premises and logical rules. A deduction is valid only if the rules of deductive logic are followed. Now, consider the following. This deduction seems clearly valid:

(1) P1: If something is an ohun then it is blue

P2: The book is an ohun

C: The book is blue

From *((Ox→Bx) & Ox)* one is entitled to conclude *Bx,* and hence a person would be rational in concluding from the deduction that the book is blue. (Given the sufficient logical abilities) it would be irrational for a person to claim that the deduction is not valid. Now, consider the following. One day Indira thinks the following deduction to herself:

(2) P1: If something is an ohun$_{(A)}$ then it is blue

P2: The book is an ohun$_{(B)}$

C: The book is blue

Indira, being a competent logician and a rational individual, thinks that this is not a valid deduction[54]. It does not follow that an object that is not blue is in fact blue, so her thinking that the deduction is invalid is rational. However, keep in mind that the label *ohun$_{(A)}$* and *ohun$_{(B)}$* is just my notational tool: it does not reflect a difference in concept, only a difference in reference. If cognitive significance turns on the vehicles of content and such entities are individuated by their origins, the two concept tokens should play the same cognitive role, since they have the same origin: "originalists reject the view that cognitive differences need to be explained by semantic differences […] Individuating concepts in a non-semantic way shows that the explanation does not rely on semantic properties of the concepts themselves" (Sainsbury and Tye 2012, 40). Reporting Indira's thoughts purely in terms of meaning vehicles would give the following deduction:

P1: If something is an ohun then it is blue

P2: The book is an ohun

C: The book is blue

But this is the same as (1), and this seems clearly valid. If cognitive significance turns on the vehicles of content and such entities are individuated by their origins, when entertained in thought, (1) and (2) should play the same cognitive role since the arguments share every concept and are structured the same way. However, (2) contains important information – information about the reference – that (1) does not contain. Hence, if concepts are individuated by their origins, Indira being rational in rejecting the deduction as invalid cannot be explained in terms of meaning vehicles; rather, the explanation relies directly on the content of her thoughts.

Now, consider a further case. The following deduction seems invalid:

[54] Hence, Indira only *considers*, without endorsing, this deduction.

(3) P1: Something is an ohun if and only if it is blue

P2: The book is not an ohun

C: The book is blue

One is not entitled to move from *((Ox↔Bx) & ¬Ox)* to *Bx*. One can only infer one side of a biconditional if given the other side. In this case, at the syntactic level, one is not given one of the sides in the biconditional, but rather the negation of the left hand side. (Given the sufficient logical abilities) it seems that one would be irrational to deem this deduction valid. Now consider the following. One day Indira performs the following deduction in thought:

(4) P1: Something is an $ohun_{(A)}$ if and only if it is blue

P2: The book is not an $ohun_{(B)}$

C: The book is blue

We have the implicit premise (P3) $ohun_{(B)} = \neg ohun_{(A)}$, regarding the content of the concept tokens. In this case, since we know that the use of $ohun_{(B)}$ is a negation of the use of $ohun_{(A)}$, we can use the double negation rule of classical logic and infer from $\neg\neg ohun_{(A)}$ that $ohun_{(A)}$. Hence, we are given the left hand side of the biconditional of (P1) from (P2) together with the rule of double negation. We are then entitled to move from *((Ox↔Bx) & Ox)* to *Bx*. Hence, Indira is to be considered rational when thinking that this deduction is valid. However, when reporting Indira's beliefs purely in terms of meaning vehicles the deduction would look like this:

P1: Something is an ohun if and only if it is blue

P2: The book is not an ohun

C: The book is blue

But this is the same as (3), and this seems clearly invalid. Note that we cannot use the rule of double negation on (3) because there is no implicit premise that allows us to take this move, since the deduction is purely syntactic, and at the level of syntax $ohun_{(A)} = ohun_{(B)}$. If cognitive significance turns on the vehicles of content and such entities are individuated by

their origins, when entertained in thought, (3) and (4) should play the same cognitive role since the arguments share every concept and are structured the same way. However, (4) contains important information – information about the reference – that (3) does not contain. Hence, Indira being rational in thinking that the argument is valid cannot be explained in terms of meaning vehicles. Rather, the explanation depends directly on the semantic features of her thought.

Sainsbury and Tye hold that "both being contradictory and being valid essentially depend on how many concepts are involved" (Sainsbury and Tye (forthcoming), 3-4). Whether this statement is true depends on what one takes a concept to be. I have shown that if concepts are taken to be mere meaning vehicles, individuated in terms of their origins, this statement does not hold. (1) and (2) contain the same meaning vehicles structured the same way, but (1) is valid whereas (2) is not. Likewise, (3) and (4) contain the same meaning vehicles structured the same way, but (3) is invalid and (4) is valid. In deductive reasoning a criterion for validity is that whenever there are equivocal terms in the premises, each token of the term must be interpreted the same way. The same holds when the deduction is entertained in thought; equivocal concepts must be interpreted in the same way in order for the deduction to be valid.[55] For (1) to be valid the two occurrences of *ohun* must have the same content. If the content of the concepts tokens are distinct, however, someone may be rational in rejecting the deduction as invalid. Hence, an individual's being rational or irrational when accepting (2) turns on the content of the concepts involved in the deduction rather than the vehicles of content individuated by their origin. If Indira rejects (2) she is rational since the argument is invalid. However, there is nothing in the formulation of (1) that accounts for her being rational in her rejecting the argument. It is only when we know the second formulation that we can explain her being rational in rejecting the deduction, but the only difference between (1) and (2) is semantic. Likewise, if Indira were to accept (2) she would be irrational. However, this cannot be explained in terms of vehicles since (1) seems valid. The explanation of her being rational or irrational, then, depends directly on semantic features rather than on the vehicles of content.

[55] For a similar point concerning the importance of this to deductions involving demonstratives, see Campbell (2002, ch. 5). Campbell considers arguments of the form '(1) That is F. (2) That is G. Therefore (3) That is F and G' and, as Mole reads him, claims that they "depend for their validity on there being no possibility of equivocating on the meaning of 'that', as it occurs in the two separate premises. Such arguments can only figure in rationally entitling reasoning so long as there is a single fixing of the referent of 'that' in both premises." Campbell's claim is that the reasoner's avoiding such equivocation relies on her attention to the referents, and that in this way the role of attention is "analogous to the role played by a Fregean Sense" (Mole 2013).

Further, Indira would be rational in accepting (4) as valid. However, if we only look at the vehicles as in (3), the argument seems invalid and she would have to be counted irrational if accepting it. Further, if Indira rejects (4) she would be irrational, but if we only consider the vehicles she would be counted rational when rejecting the deduction. If taking $ohun_{(A)}$ and $ohun_{(B)}$ to be of the same type, the only difference between (3) and (4) is due to semantics. Hence, the explanation of Indira being rational or irrational is semantic rather than syntactic. The upshot of this is that if taking cognitive significance to essentially depend on the vehicles of content rather than the content expressed, Sainsbury and Tye fail to explain certain cases of rationality since they allow for the number of concepts involved in thinking come apart from the number of possible contents expressed by such entities. According to originalism, thinkers are only related to content via thoughts, but a result of their theory is that thoughts of the same type may be related to distinct content. In order to explain the case just outlined it seems that one must allow for cognitive significance to rely directly on content rather than vehicles. But Sainsbury and Tye deny that thinking involves standing in a relation to any kind of content, and hence the theory fails to account for Indira being rational. Since originalism cannot explain certain cases of rationality, the theory fails as a general account of the nature of concepts and thoughts.

When setting forth the thought experiment I've worked on the assumption that Sainsbury and Tye hold that every instance of cognitive significance can be explained by appeal to vehicles of content, rather than the content expressed by such entities. It may, however, be objected that Sainsbury and Tye are not committed to such a strong claim. Even if they hold that the seven puzzles can be explained by appeal to vehicles alone it may be that they hold that other cases of cognitive significance can only be explained by taking the reference into account. If this is the case, the thought experiment, the way it was just presented, does not serve as a criticism of the theory, but rather as a qualification of the view, illustrating that originalists must in fact take reference into account when explaining certain cases of cognitive significance. However, by making some adjustments to the thought experiment I will, in 5.3, show that even if allowing the reference to be part of the explanation, there are still certain cases of cognitive significance that cannot be explained within Sainsbury and Tye's framework. Importantly, Sainsbury and Tye's framework commits them to explaining cognitive significance in terms of vehicles and reference alone, both individuated historically.

5.3 The Thought Experiment Expanded

Now, consider the following. After having lived with group B for some years, Indira travels on to a new location, and settles down at this location for 40 years. During this time, the concepts *ohun*$_{(A)}$ and *ohun*$_{(B)}$ have a new change in reference: Gradually, both group A and group B come to use the concept to pick out anything that is an object (just like the original concept *ohun*). The concepts are still the same type, since the change in reference is stipulated to be gradual and with every use involving deference to earlier uses. The situation, then, is this: Unbeknownst to Indira, *ohun*$_{(A)}$ and *ohun*$_{(B)}$ now have the same reference, and both uses of the concept have a different reference than she thinks.

After having been away for 40 years, Indira returns to group B. Upon her arrival, nobody informs Indira of the change in reference of the concept *ohun* and for a long time Indira only encounters uses of the concept that do not reveal to her that there has been a change in reference. She hears people say such things as 'many ohuns are pretty' and 'my ohuns occupy all the space in my living room', but these kinds of sentences make sense to Indira even if she takes the utterer to mean only objects that are blue or only objects that are not blue, so she never questions the reference of *ohun.* Importantly, Indira still defers to the other people in her language community when using the concept *ohun*, so Sainsbury and Tye must say that Indira uses the same concept as does the others even if she makes a mistake about the reference of the public concept: again, "Concept possession is consistent with all sorts of mistakes and misunderstandings about the concept's subject matter" (Sainsbury and Tye 2012, 55). Furthermore, Sainsbury and Tye, recall, contrast indexical concepts with public concepts such as *Paderewski,* and claim the following about the latter: "It's not up to individual users to settle anything about the nature or semantics of that concept" (Ibid., 52). Since Indira uses the public concept *ohun,* the content of her tokened concept should be the same as the public use, on this view.

After her return, Indira is still able to form beliefs such as *sapphires are ohuns*$_{(A)}$ and *sapphires are ohuns*$_{(B)}$*,* and also *sapphires are not ohuns*$_{(A)}$ and *sapphires are not ohun*$_{(B)}$*.* This time, however, since the reference of *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are the same – anything that is an objects – only the first set of beliefs is true. Still, to Indira it seems as though the thoughts *sapphires are ohuns*$_{(A)}$ and *sapphires are ohuns*$_{(B)}$ are contradictory, since she believes *ohun*$_{(A)}$ to refer to objects that are blue and *ohun*$_{(B)}$ to refer to objects that are not blue. Hence, if Indira were to form both beliefs she would be irrational. Likewise, just like before the second

change in reference, Indira would be rational in forming both the beliefs *Sapphires are ohuns$_{(A)}$* and *sapphires are not ohuns$_{(B)}$.* However, since after the second change in reference the two thoughts share the same concepts with the same reference, except from the negation, the latter must be a denial of the first, according to Sainsbury and Tye. Hence, Indira should be counted irrational for forming both beliefs, but *ex hypothesi* she is not. This shows that, somehow, Indira's beliefs about the reference of *ohun*$_{(A)}$ and *ohun*$_{(B)}$ must be relevant to the cognitive role of her thoughts. The question, then, is in what sense Indira's beliefs about the reference is relevant to the explanation of cognitive significance. There are three possibilities available to Sainsbury and Tye: (a) they could say that Indira uses only one concept with the same reference. If this is the case, they must appeal to something else than their core framework in order to explain Indira being rational; more specifically Indira being wrong about the content of her beliefs; (b) they could hold that Indira's beliefs are the same at the level of vehicles but that they differ in content due to Indira's beliefs about the reference of the concept; or (c) they could hold that Indira's beliefs are essential for the individuation of the concept *ohun,* and thus say that her thoughts contain distinct concepts. I will show that none of these routes are viable: Route (a) fails, and route (b) and (c) are not available to Sainsbury and Tye's originalism.

The first possibility is to explain the difference in cognitive significance of the two thoughts in a similar way to how Sainsbury and Tye explain the puzzle of *Paderewski.* The case just outlined bears similarity to the puzzle of Paderewski, but it also differs from Kripke's puzzle in some important respects. First, while everyone who takes concepts to be public should agree that Peter only uses one concept *Paderewski,* not everyone who takes concepts to be public have to agree that Indira only uses one concept *ohun.* The reason why Sainsbury and Tye must say that Indira only uses one concept *ohun,* is that they take concepts to be individuated by their origins, and the fact that they hold that a difference in content does not necessitate a difference in concepts; theories who reject these claims would have no problem explaining Indira being rational. Further, and importantly, unlike Peter, Indira is not wrong about the number of concepts she uses. Hence, Indira's having a false second order belief is not due to her being wrong about the number of concepts deployed in thought – the knows the relevant concepts to have the same origin. This shows that their solution to the Paderewski case is not applicable here. Instead, Sainsbury and Tye may say that Indira having a false second order belief about the content of her first order beliefs. This route is available to Sainsbury and Tye due to them holding that individuals can be wrong about the content of

their concepts as well as the vehicles (see footnote 41 on page 43 in this thesis). The question we wish to explain is this: How can it be that Indira can have contradictory beliefs[56] and still be counted rational? The suggestion is that one could try appealing to Indira having a false belief about the content of her beliefs *sapphires are not ohuns*$_{(B)}$ and *sapphires are ohuns*$_{(A)}$, in order to explain her being rational. The explanation would then look like this: When forming the belief *sapphires are not ohuns*$_{(B)}$, Indira is rational since she believes that she does not already have a belief with contradictory content to the content expressed by her new thought (to be sure, let's say one of Indira's friends from group B, who has previously heard Indira say that she believes that *sapphires are ohuns,* asks her if she is sure she doesn't already have a belief with contradictory content; Indira would give a negative answer). However, if someone from group A was to ask her the same question, she would tell them that she has a belief that expresses a contradictory content to the thought *sapphires are not ohuns*$_{(A)}$ (namely the belief *sapphires are ohuns*$_{(A)}$). After the second change in reference, however, the content of Indira's belief *sapphires are not ohuns*$_{(A)}$ and *sapphires are not ohuns*$_{(B)}$, are exactly the same: the set of worlds in which sapphires are not objects. Hence the two thoughts share the same structure and express the same content. How can it be then, that Indira, being rational, can have both the belief that she does not have a belief with contradictory content to the content expressed by the thought *sapphires are ohuns* and at the same time think that she *does* have a belief with contradictory content to the content expressed by the thought *sapphires are ohuns*?

One could try to appeal to Indira being wrong about the content of these new contradictory beliefs. To be clear, Indira's two new contradictory beliefs are the following: Let (Q1) be the belief that she does <u>not</u> have a belief with contradictory content to the content expressed by the thought *sapphires are not ohuns*[57]. Let (Q2) be the belief that she *does* have a belief with contradictory content to the content expressed by the thought *sapphires are not ohuns.* Given that the two thoughts contain the same public concepts (except for the negation) with the same public content, these thoughts seem clearly contradictory. We need to explain how Indira could be rational in forming (Q2) after having formed (Q1). One could try appealing to

[56] Note that since on this version of the thought experiment, Sainsbury and Tye do not need to appeal to vehicles alone in order to explain cognitive significance, two thoughts are contradictory only if they share a complete tree structure except from one negating the other *and* if they also express contradictory content (i.e. they can never both be true since the sets of worlds in which they are true do not overlap).

[57] The notation 'ohun$_{(A)}$' and 'ohun$_{(B)}$', recall, was intruduced as a notational tool in order to reflect a difference in reference. Importantly, the notations do not reflect an individual's own conception of – or beliefs about the reference. Since, after the second change in reference there are no longer two distinct public contents, I will not use the notation further in this chapter; I will simply use *ohun.*

Indira having a false belief that she does not already have a belief with contradictory content to (Q2). However, if asked by someone from group B if she already has a belief expressing contradictory content to (Q2), she would say that she *does* have such a belief.[58] If explaining these contradictory beliefs in terms of Indira having false beliefs about the content of these, we would end up with even more sets of higher order contradictory beliefs about the content of her thoughts. As I have argued, one cannot explain someone being rational in having seemingly contradictory first order beliefs by appeal to false higher order beliefs if one ends up having to attribute even more sets of contradictory beliefs the further up the chain. Hence, this route fails to explain Indira's being rational.

The second possibility is for the originalist to say that the content of Indira's concept tokens are not the same as in the public use. Rather, the contents of Indira's concept tokens are the same as before she left the group more than 40 years ago. This case, then, resembles the case of *Madagascar,* addressed in 2.4. Marco Polo, recall, used the same concept *Madagascar* as the locals, but with a different reference despite intending to use the concept the same way as the locals. On this picture, Indira uses the concept *ohun* with a different content than others in her community despite of her deferring to the others' uses. In 2.4, however, I argued that even if Sainsbury and Tye allow for these kinds of situations, it is mysterious how they are to be explained. Since, according to Sainsbury and Tye, the content of concepts are determined by a deference to the public use, and does in no way depend on any such thing as the thinker's cognitive content, it seems that these kinds of scenarios shouldn't be possible. If Sainsbury and Tye wish to explain Indira's being rational in terms of her concepts having distinct reference, then, they must abandon the claim that the content of concepts are to be individuated historically. But if they take the reference to be determined by a thinker's cognitive content (in the sense of associated descriptions) we are no longer operating within a Kripkean/Millian framework. One of the main motivations behind originalism is to explain Fregean data within a Millian framework, so this route is not viable to the originalist. Unless originalists can give an account of change in reference without appealing to a thinker's own associated descriptions – and at the present moment it is hard to see how they can provide such an account – this route is not available to them.

[58] To be clear, the reason why she would answer differently when asked by someone in group A and group B is because she has a false belief about the content of the public concept *ohun*: Just like Peter in the Paderewski case was assuming his two friends to be asking him about distinct individuals when enquiring him about Paderewski, Indira wrongly takes people in group A and group B to ask question involving concept tokens with contradictory content when asking her question involving the concept *ohun.*

The last possibility is to say that Indira's beliefs about the reference – i.e. her conception of it – are essential to the individuation of the vehicles. On this route cognitive content (in the sense of associated beliefs) is at least partly essential to the individuation of concepts. One could then say that, to Indira, *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are distinct concepts due to the difference in associated content, and then explain the difference in cognitive significance in terms of a difference in vehicles. This, however, violates key assumptions of originalism, according to which the vehicles are to be individuated by the origins alone, and also the claim that cognitive significance can be explained without appeal to anything beyond vehicles and reference; especially without appeal to cognitive content. Hence, this route is not available to the originalist.

Based on the account given in *Seven Puzzles of Thought,* I see no further explanations available to Sainsbury and Tye. This shows that there are cases in which cognitive significance cannot be explained by appeal to vehicles of content and reference alone, if taking such entities to be public and individuated historically. Hence, Sainsbury and Tye's formulation of originalism fails as a general account of concepts and thoughts since it fails to explain certain cases of cognitive significance.

5.4 Chapter Summary

In this chapter I have argued that originalism about concepts and thoughts is insufficient for giving a general account of rationality. I started by showing that, if allowing concepts to change their reference, it is possible for two concept tokens of the same type to express contradictory content. If one holds that cognitive significance is to be explained in terms of vehicles alone, two thoughts that share all compositional concepts and syntactic structure should play the same cognitive role. However, I have presented a case in which it seems that two such thoughts play different roles in cognition. For instance, (in the first formulation of the thought experiment) the thought *sapphires are ohuns*$_{(A)}$ is true while the thought *sapphires are ohuns*$_{(B)}$ is false, so they should play different cognitive roles, but if one, as the originalist does, takes cognitive significance to depend solely on public concepts - construed as vehicles of content - individuated by their origins, one must say that the thoughts are of the same type and thus the thoughts should play the same role in cognition. Someone forming both of the beliefs while knowing them to have distinct truth conditions would be irrational, but since the

thoughts (meaning vehicles) are the same, according to originalism, originalists would have to appeal to something else (plausibly semantic features) in order to explain rationality. Further, the thought *ohuns$_{(A)}$ are ohuns$_{(B)}$* is an identity statement containing concept tokens of the same type, and it should therefor be trivial. However, the thought is *not* trivial, and the level of informativeness can only be explained by appeal to the content of the thought.

It seems to follow from Sainsbury and Tye's account that a subject who knowingly equivocates when performing a deduction in thought may nonetheless be counted rational.[59] If someone can use concepts of the same type to express different contents, a deduction (involving those concepts) that would be valid at the level of syntax may be invalid when taking the content into account. Hence, if taking the cognitive processing involved in the deduction to essentially depend on the number of concepts (vehicles of content), this is insufficient for explaining why someone is rational or irrational in accepting the deductions I've presented as valid. The conclusion to draw from the thought experiment is this: If one allows the individuation of the vehicles of content to come apart from reference fixing, one cannot give a general account of rationality in terms of vehicles of content alone. In certain cases originalists must appeal directly to the content of concepts and thoughts rather than the vehicles of content.

In the final part of this chapter, I argued that even if Sainsbury and Tye agree that certain cases of cognitive significance cannot be explained without appealing directly to the content of thoughts rather than the vehicles, their theory still fails at giving a general account of rationality. By making some adjustments to the original thought experiment, I argued that Sainsbury and Tye cannot explain how Indira can be rational in forming thoughts that to her seem consistent, but that in reality are contradictory due to the public use of the constituent concepts. In both 5.2 and 5.3, the problems posed for the theory was due to their claim that a change in content does not necessitate a change in vehicles. If Sainsbury and Tye did not allow for this, they could explain both versions of the thought experiment by appeal to a difference in vehicles. However, since the key originalist claim is that concepts are individuated by their origins and not by semantic or epistemic properties, it seems that the originalist framework cannot allow them to do this. Hence, if my argumentation is successful,

[59] Note that some hold that one may in some cases be rational in engaging in equivocal (and hence invalid) reasoning. But in these cases, the subject equivocates unknowingly (due to slow switching) (cf. Gerken 2013, Recanati 2012)

the thought experiment shows that Sainsbury and Tye's originalist theory fails as a general theory of concepts and thoughts, since it fails to explain certain cases of rationality.

In the next chapter I will look into possible solutions to my thought experiment on behalf of Sainsbury and Tye. More specifically, I will consider the possibility of them denying that $ohun_{(A)}$ is the same concept as $ohun_{(B)}$. If the concepts are not of the same type, the thought experiment is not effective. I will, however, argue that Sainsbury and Tye's originalist account as stated in *Seven Puzzles of Thought* is in fact committed to saying that the concepts are of the same type. If the originalists cannot solve the puzzle posed for the theory in this chapter, the theory is rendered unattractive as a general account of concepts and thoughts.

Chapter 6

Possible Solutions

In chapter 4 I showed that Sainsbury and Tye's originalism fails to solve three of the puzzles it is advanced to solve. In chapter 5 I argued, on the basis of a thought experiment, that their theory encounters further problems in accounting for certain cases of rationality. If my take on Sainsbury and Tye's account is correct, the theory fails as a general theory of concepts and thoughts. In this chapter, I lay out what I take to be the most prominent objections to – and possible solutions of my thought experiment from chapter 5 on behalf of originalists.[60] In particular, I look into the possibility of originalists denying that the concepts involved in the thought experiment, *ohun*$_{(A)}$ and *ohun*$_{(B)}$, really are concepts of the same type. If originalists can avoid claiming that the tokened concepts are of the same type, my thought experiment is ineffective. I will, however, argue that Sainsbury and Tye are committed to saying that *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are of the same type.

6.1 Giving an Alternative Explanation of *Meat*

A possible solution to my thought experiment on behalf of Sainsbury and Tye may be to deny that *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are the same concept. If they are not, the difference in cognitive significance of thoughts such as *sapphires are ohuns*$_{(A)}$ and *sapphires are ohuns*$_{(B)}$ can be explained by appeal to the meaning vehicles. Further, the deductions addressed in 5.2 could be explained by appeal to syntax; deduction (1) would be the same as deduction (2), and deduction (3) would be the same as (4). If this is the case, Indira's being rational can be

[60] More specifically, originalists who endorse Sainsbury and Tye's version of the theory (they make certain claims that not all originalists are committed to.)

explained in terms of vehicles of content; also in the second formulation of the thought experiment. I will, however, show why this is not an option for proponents of originalism as stated in *Seven Puzzles of Thought.*

Even though Sainsbury and Tye's theory of how concepts acquire and maintain their content resembles Kripke's causal theory of reference, they agree with Evans that in some cases, the content of concepts may change while the concepts remain the same. In the case of *meat* Sainsbury and Tye hold that the concept has remained the same but with a change in representational content; the thought *squash is meat* tokened by someone in the fifteenth century contains the same conceptual structure as when tokened by someone today, but the truth conditions of the two thought tokens are distinct. I have argued that the case of *ohun* is analogous to the case of *meat.* When discussing the case of *meat,* however, Sainsbury and Tye say that "another option is to say that a new concept, expressed by a word spelled and pronounced the same way, was introduced at some point, and each of the two concepts have retained their original and distinct contents" (Sainsbury and Tye 2012, 46). This might seem like a viable route for the originalists, but I will argue that this is not really an alternative for originalists that agree with Sainsbury and Tye's conditions for being a non-originating use. These conditions are the following:

1) The use involves deference to other uses, by the same subject or other subjects.
2) The use involves informational accumulation from other uses, by the same subject or other subjects. (Sainsbury & Tye 2011, 2)

The history of the concept *meat* is stipulated to satisfy both conditions: At no point in history did any user of *meat* not defer to other uses of the concept, and every individual in the chain of deference accumulated information about the concept from earlier uses by herself and others in her language community. The change in reference was not due to any intentional re-fixing of the reference, but rather tiny unnoticeable deviations of standard use that are stipulated to be too small to make the use count as an originating use.[61] However, it might be possible for the originalists to say that these unnoticeable deviations did in fact constitute an introduction of the concept *meat* that we use today. If this is the case, our concept *meat* is

[61] See 2.4 for criticism of the claim that a concept may change its reference despite the user deferring to other uses.

distinct from the concept *meat* in the fifteenth century.[62] This move, however, is not a good option for originalists like Sainsbury and Tye that take concepts to be public. If the error regarding the reference of *meat* happened gradually and with no deviation being more serious than others, we are not given the equipment to decide where in the chain a new concept originated. If two deviations are equally serious and one of them constitutes the origin of a new concept, so should the other. And if we stipulate the total number of equally serious deviations in the history of the concept meat to be one thousand, say, it seems like if one of these is a deviation serious enough for the use to count as a originating use, so should the others. But then, instead of having two concepts *meat* we would have a thousand and one distinct concepts expressed by the same word. If all of these concepts survived through time, there is not one concept *meat* today but one thousand! But this seems implausible. Further, if the tiny deviations in the history of *meat* constitute the introduction of new concepts, the same might be the case with several of the public concepts you and I are taken to share on Sainsbury and Tye's originalist framework. For instance, you might make a tiny deviation from standard use when using the concept *chair*, and so might I. If the deviations in the history of *meat* were serious enough to introduce a new concept, so might our deviations. Even though I defer to other people in my language community when using the concept *chair* it may be that I've made a small unnoticeable error regarding the reference of the concept (for instance, I might think that a broad chair is a sofa rather than a chair). Further, the same might be the case with our concepts *table, cat, tea* and so on. It seems plausible that such tiny errors regarding the reference of a concept occur all the time: for instance, you and I could both be confused whether broad chairs are chairs or sofas; you might think they are chairs, whereas I think they are sofas. But then, if deviation from standard use is sufficient for the introduction of new concepts, it seems like you and I use distinct concepts from each other. Also, it might be the case that none of us qualifies as users of the public concept. Such cases are so common, that it seems to render the notion of public concepts that Sainsbury and Tye use with little or no explanatory value. If every tiny deviation were an instance of an introduction of a new concept most uses would be an originating use of a concept. Then deference would loose its utility, since one would have to grasp the reference of a concept completely in order to count as a user of a given public concept. This is too strong a demand,

[62] To be clear, Sainsbury and Tye would not want to say that the kind of errors found in the history of *meat* are sufficient for the introduction of a new concept; the sameness in the vehicles of content is ensured by deference to earlier uses, and concept possession is "consistent with all sorts of mistakes and misunderstandings about the concept's subject matter" (Sainsbury and Tye 2012, 55). I present this possibility just in order to block this maneuver.

and Sainsbury and Tye do not think that grasping the content of a concept is necessary for acquiring that concept. Hence, as long as originalists take concepts to be public, this is not a viable route.

Another possible route is to deny that every deviation throughout the history of the concept *meat* constituted an introduction of a new concept, but that one did. However, this renders us with a situation resembling a sorites paradox: If one unnoticeable deviation does not constitute an originating use, neither should the next equally unserious deviation, nor should the third deviation, and so on. It then seems like no amount of such deviations can constitute an originating use, but on this alternative route suggested (but not taken) by Sainsbury and Tye, somewhere along the chain of deference a new concept should come into existence. Originalists hold that there is a unique point in history for each concept at which they come into existence. Hence, since a concept can only be used intentionally for the first time once, the point in history at which a concept come into existence cannot be fuzzy. The problem in the case of *meat* then, is that we are not given the equipment needed to decide at what point in the chain of deference a new concept come into existence. Hence, at best, this route leads to a paradox; if no single deviation from standard use is sufficient for the introduction of a new concept *meat* (and for the originalist there has to be exactly *one* such historical point at which the new concept was introduced), how can such an introduction take place when there's a gradual drift? Originalism as stated in *Seven Puzzles of Thought* fails to give an account of how this can be the case. Given central claims of their originalist theory, then, it seems that Sainsbury and Tye are committed to treating the case of *meat* as a case of a concept staying the same but with a change in reference. In my thought experiment I took the history of *ohun* to be analogous to the history of *meat.* There is, however, one obvious difference between the two cases; in the case of *ohun* the same concept came to express contradictory contents at the same time in history, whereas the reference change in *meat* was so that the concept only had one content at a time. I will go on to discuss whether the originalist can say that the history of *meat* is a case in which a concept changes its content but remain the same type, whereas the case of *ohun* is a case of new concepts coming into being.

6.2 The Case of *Meat* Being Different from the Case of *Ohun*

One possible route might be for the originalist to say that the history of *ohun* is an instance of conceptual fission: at some point in history, the original concept *ohun* fissioned into two concepts; *ohun(A)* and *ohun(B)*[63]. If this is the case, neither group (A) nor group (B) use the same concept as the original concept *ohun* that originated in 1300. *Ohun(A)* originated when someone intentionally used the concept to pick out only objects that are blue for the first time, while *ohun(B)* originated when someone first used that concept to pick out only objects that are not blue. If this is the case, *ohun(A)* and *ohun(B)* are distinct concepts since they have distinct origins. This scenario, however, is blocked by the stipulations made in the thought experiment. I stipulated that every use of the concept *ohun* through the history involved deference to earlier uses and information accumulation from others in one's language community, which is completely in line with originalism. Earlier in this chapter we saw that Sainsbury and Tye take this to be sufficient for the use to count as a non-originating use. But in 2.3 we saw that conceptual fission is a matter of new concepts originating. An originating use does not involve any deference to earlier uses, but rather the intention to introduce a not already existing concept. At no point during the history of *ohun* did anybody intend to introduce a new concept lexically identical to the already existing concept. Hence, since the people in group A and group B all intended to use the same concept *ohun* as others in their language community, it cannot be an instance of conceptual fission, since fission involves an intention to introduce a new concept.

I will now look into a further suggestion as to how originalists can deny that $ohun_{(A)}$ and $ohun_{(B)}$ are the same concept, and argue that this route also fails. In *Seven Puzzles of Thought* Sainsbury and Tye describe the structure of the chains of deference having the same origin as a tree structure where all concept tokens belonging to one of the branches are of the same type: "For originalists, what makes two uses uses of the same concepts is their belonging to a single use-tree" (Sainsbury and Tye 2012, 88). Maybe it is possible for Sainsbury and Tye to reject the view that $ohun_{(A)}$ and $ohun_{(B)}$ are the same concept if modifying this claim. Perhaps they could say that concepts are of the same type if and only if they belong to the same branch of such a tree of deference. That is to say, concepts are of the same type if and only if they have the same origin *and* belong to the same chain of deference. I will explore this possibility and show why it is an unattractive choice for originalists.

[63] Where this time the notation (A) and (B) signals a difference in type of concept

If one takes concepts to be of the same type only if they belong to the same chain of deference we could draw the following picture of Indira's use of *ohun*. When using $ohun_{(A)}$, Indira defers to people in group A. When using $ohun_{(B)}$, she defers to people in group B. Since the chain of deference is distinct the concepts are distinct. In the case of *meat,* one could then argue that there is only one chain of deference, and this assures a conservation of the original concept. While this has the welcomed consequence that $ohun_{(A)}$ and $ohun_{(B)}$ are distinct concepts, it also follows that all concepts used by group A are distinct from the concepts used by group B, since the two groups always defer to different groups of people. Sainsbury and Tye cannot say this since they hold that concepts are public and sharable. If group A and group B do not count as users of the same concepts, as a result of them deferring to different groups of people, then it would seem as thought people in the US and people in the UK also use distinct concepts from each other. On this picture people in the US defer to other speakers in their community when using a concept, while people in the UK defer to the use of their locals when using a concept, and since they defer to different groups of people they do not count as users of the same concepts.[64] Further, small children having just acquired a certain public concept may defer only to their parents when using that concept. Since small children defer to different uses (different sets of parents) it seems that they are using distinct concepts, on this picture. But even so, they are using the same concept as their parents, and the parents may defer to a larger group of people. If the parents in one community use the same concept, then, so should their children; identity is transitive, so if use (a) = use (b) and use (b) = use (c) then use (a) = use (c). However, if one must share a complete deferential chain in order to use the same concept, it seems as though the children cannot use the same concepts as each other. The same point holds for the case of *ohun*: Everyone in group A defers to earlier uses of the concept all the way back to the point in history at which *ohun* was first used to pick out anything that is an object. Hence the originalist must say that people in group A use the same concept as did people in the 14th century. The same holds for group B: Everyone in group B defers to earlier uses all the way back to the point in history at which *ohun* was first introduced. Hence, the originalist must say that they also use the same concept as did people in the 14th century. If group A uses the same concept as the original concept and group B also uses the same concept as the original concept, transitivity has it that the two groups must use

[64] A further problem with this route is how to determine where to draw the line between different groups. What are the necessary and sufficient conditions for someone being part of one group rather than another? And what should be said about individuals who seem to belong to two different groups, e.g. individuals who commute between the US and the UK? It seems strange to say that such individuals use different sets of concepts depending on their current location.

the same concept as each other. But if holding that concepts must share a complete chain of deference in order to be of the same type, the concepts cannot be of the same type. Hence, this view leads to paradoxes, and it is unlikely that the originalists would want to say that two concept tokens are of the same kind only if they share a complete chain of deference.

It seems, then, that Sainsbury and Tye must agree that *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are the same concept. At least, we are given no resources in *Seven Puzzles of Thought* to argue otherwise.[65] If they must agree that the concepts are of the same type, thoughts such as *sapphires are ohuns*$_{(A)}$ and *sapphires are not ohuns*$_{(B)}$ are in fact contradictory on their view. Further, the thoughts *sapphires are ohuns*$_{(A)}$ and *sapphires are ohuns*$_{(B)}$ must be of the same type, since they share an entire conceptual structure, and hence they should play the same role in cognition, according to originalism. Hence my argument in chapter 5 is effective.

6.3 Chapter Summary

In this chapter I have presented what I take to be the most prominent answers and objections to my thought experiment on behalf of the originalists, and argued that these replies fail. More specifically, I have argued that Sainsbury and Tye cannot deny that *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are distinct concepts. In chapter 5 I stipulated the history of *ohun*$_{(A)}$ and *ohun*$_{(B)}$ to be analogous to the history of *meat.* In this chapter I looked into the possibility of originalists saying that, in the case of *meat,* a new concept originated somewhere in the history. I showed that Sainsbury and Tye are in fact committed to saying that the history of *meat* is a case of a concept staying the same but with a change in reference. Since the history of *ohun* is stipulated to be analogous to the history of *meat,* originalists must say that *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are of the same type. One might, however, argue that the case outlined in chapter 5 is not really analogous to

[65] If, nonetheless, it turns out that the originalist could deny that *ohun*$_{(A)}$ and *ohun*$_{(B)}$ are of the same type due to them expressing contradictory content at the same time in history, it is possible to construct a thought experiment similar to the one I laid out in chapter 5, but in which there are not two public uses of the same concept at the same time in history. Consider a time-traveller who travels back in time to the 15th century. When using the concept *meat,* deferring to the people he meets back in time, he uses the same concept as he did while still in the 21st century. However, he intends to use the concept the same way as others in his new community, so for him the content of *meat* includes anything edible, not just animal flesh. When deferring to people in the 15th century, then, his thought *squash is meat* is true. When talking with his family in the 21st century on his time-travel phone, however, he intends to use the concept the way *they* use it. Hence, when deferring to his family the thought *squash is meat* is false. Since one of the thought tokens are true and the other false, they should play distinct roles in cognition. They do however share a complete tree-structure, according to originalists. Hence, the thoughts playing different roles in cognition cannot be explained by appeal to the vehicles of content.

the case of *meat* since the result is the same concept having distinct content at the same time in history. One might, then, take the case of *ohun* to be a case of conceptual fission. This cannot be the case, however, since conceptual fission requires someone intending to introduce a new concept and not defer to earlier uses. The final possibility addressed was for the originalist to hold that concepts must share the exact same chain of deference in order to be of the same type. I argued that this route leads to paradoxes; if each token in every branch of a deference tree are the same as the original concept, by transitivity, they should be the same as each other, but if concepts must share a complete chain of deference in order to be the same concept, transitivity fails. Hence, this route is rendered nonviable.

Since all of the replies to my thought experiment addressed in this chapter fails, my criticism of originalism stands non-refuted. In chapter 4 I argued that originalism fails to explain three of the puzzles of thought – puzzles the solutions of which are the raison d'être of the theory. The fact that originalism fails to solve these puzzles, together with the problems the theory encounters with my thought experiment, renders originalism unattractive as a general theory of cognitive significance and rationality. In the next chapter I will suggest an alternative to originalism. The theory I will take a look at is François Recanati's theory of mental files, presented in his (2012). Recanati's theory is in many respects similar to originalism; for instance, Recanati agrees with Sainsbury and Tye that one can employ a non-semantic understanding of vehicles of content in order to explain cognitive significance. But, as will become clear in the next chapter, the two theories also differ in important respects. I will show that Recanati's account of mental files is the better alternative, since it is better suited to solve the classic puzzles; in particular the puzzles addressed in chapter 4 of this thesis. I also show that Recanati avoids the kind of problems posed by my thought experiment. Note that the chapter is not intended as a general argument in favour of Recanati's theory – giving such an argument is beyond the scope of this thesis. The purpose of presenting Recanati's theory is, rather, to show that there are theories of cognitive significance that are better suited than originalism to handle the problems discussed in this thesis (and one might get a suggestion as to what is missing on the originalist account).

Chapter 7

Mental Files: An Alternative to Originalism

In his *Mental Files* (2012) François Recanati puts forth a theory of singular thoughts according to which the constituents of thoughts are to be understood as *mental files*, a notion with which he claims to be able to solve some of the traditional problems in the theory of thoughts and concepts. The theory bears similarity to originalism: His notion of a mental file is analogous to Sainsbury and Tye's notion of a concept: mental files are singular terms in the language of thought. These files are, like Sainsbury and Tye's concepts, non-semantic vehicles of content. However, an important difference between the two theories is that while Sainsbury and Tye take the constituents of thoughts to be public, Recanati takes these to be individual. Further, Recanati does not hold, as do the originalists, that vehicles of content are to be individuated by their origins. In this chapter I will give a brief overview of Recanati's account and investigate whether this theory faces the same problems as originalism. If Recanati's theory of mental files can explain the different puzzles of thought and avoid the problems posed by my thought experiment, we have reasons to prefer Recanati's theory to originalism. As noted, I do not intend this to be a general defence of Recanati's theory; there might be problems with Recanati's theory that I do not address in this thesis. Here the focus will be on the seven puzzles and my own thought experiment, and whether or not Recanati's theory provides a better explanation of these than originalism. First, I give a very brief introduction to Recanati's theory, and compare it with the originalist view. I then show how the mental file framework can solve the seven puzzles of thought. Finally I, show that my thought experiment poses no problems for Recanati.

7.1 Recanati's Theory of Mental Files

Recanati offers a novel theory concerning thoughts and what it is to think about objects. At all times, individuals stand in relations to objects in their environment. Some of these relations – the acquaintance relations – provide us with knowledge about the objects we are related to: "In general there is acquaintance with an object whenever we are so related to that object that we can gain information from it, on the basis of that relation" (Recanati 2012, 20). Recanati calls these relations *epistemically rewarding* (ER) relations. According to Recanati the information gained from such relations are stored within *mental files*. The mental files are vehicles of content; they are not themselves semantic, but express semantic content.[66] All mental files correspond to an acquaintance relation and the function of the files is to store information about the object to which one is related. Recanati takes these relations to be essential to the type of the files and he takes the files to be individuated by their function to store information gained from the ER relations. For instance, when I'm perceptually acquainted with Big Ben, a mental file is created in my mind and the information gained about Big Ben is stored in this specific file. On Recanati's view, the mental file created when perceiving Big Ben is to be individuated in terms of the file's function to store information about Big Ben (which it has in virtue of standing in an appropriate ER relation to it).

There are different kinds of mental files; there are *proto-files,* which can "*only* host information gained in virtue of the ER relation to the referent" (Recanati 2012, 64). Such proto-files can only gather information from one unique ER relation. The paradigmatic case of an ER relation is perceptual acquaintance. There are, of course, other ways to gain information about objects than by standing in a direct relation to the referent itself; a large part of our knowledge and beliefs is based upon information gained from others. Files capable of containing both kinds of information are what Recanati calls *conceptual files*. These files contain both "information gained in the special way that goes with that relation […] and information not gained in this way but *concerning the same individual as information gained in that way*" (Ibid., 65-66). Such files can gather information from more than just one ER relation. I will return to this distinction in 7.2, where I show that Recanati's mental file framework is better suited than originalism to solve the seven puzzles of thought.

[66] In this respect they correspond to Sainsbury and Tye's notion of concepts.

We saw that Recanati takes the individuation of the mental files to depend on ER relations. He also takes the references of the mental files to be determined relationally: Mental files refer to whatever they gain information from through ER relations. The reference of a mental file

> is not determined by properties which the subject takes the referent to have (i.e. by information – or misinformation – *in* the file), but through the relations to on which the files are based. The reference is the entity we are acquainted with (in the appropriate way), not the entity which best 'fits' information in the file (Recanati 2012, 35).

Recanati's account, then, provides an answer to question (1) what is the content of concepts, and question (2) in virtue of what does a concept have its reference, discussed in chapter 1, without appeal to two levelled semantics. Importantly, since the ER relation both determine the type of the vehicles (mental files) and also the reference of such entities, the individuation of concepts and the fixing of reference does not come apart on this framework. This will be important for the discussion in 7.3, where I show that Recanati does not face any problems with my thought experiment.

Recanati agrees with Frege that cognitive significance cannot be explained within a classic Millian framework. He does not agree, however, that what is needed to explain this phenomenon is a further layer of semantics[67]. Recanati uses the term *cognitive content*, but importantly, to him, cognitive content is non-semantic. Cognitive content, according to Recanati, is characterized by the functional role of the files deployed in our thoughts. Hence, he rejects the descriptivist view that cognitive content (mode of presentation) is a set of associated descriptions. In contrast, Recanati takes singular thoughts to be directly about individual objects as much as they are about properties (Recanati 2012, 3-14). Originalists agree with Recanati that one does not need a semantic notion of mode of presentation in order to explain cognitive significance. Originalists do, however, hold that cognitive significance can be explained without having to attribute *any* notion of cognitive content (be it semantic or functional) to concepts and thoughts. An important difference between Recanati and Sainsbury and Tye, then, is that Recanati thinks that there is a distinction between cognitive content and referential content and that the cognitive content determines both the mental files and their reference, whereas originalists reject that there is any such thing as cognitive content. According to originalists there is no such single mechanism that account for the individuation of concepts and their reference; even though both depend on deference, they

[67] This is a further respect in which Recanati's approach is similar to Sainsbury and Tye's.

have a two levelled understanding of deference, allowing for individuation and fixing of reference to come apart.

To sum up; Recanati takes the vehicles of content (mental files) to be individuated by their cognitive contents, i.e. their function to store information about the objects to which they stand in ER relations. The ER relation also determines the reference of the files; mental files refer to the object they stand in an ER relation to. In the next section, I will apply Recanati's theory of mental files to the seven puzzles of thought addressed by Sainsbury and Tye. If Recanati can provide better solutions to the puzzles, and in particular avoid the objections I raised for the originalist account in chapter 4, we have reasons to prefer Recanati's view to originalism. Note that in *Mental Files* Recanati discusses explicitly only the case of Hesperus and Phosphorus, the case of thinking about oneself and the puzzle of empty thoughts. The solutions to the other puzzles are my proposals on behalf of Recanati.

7.2 The Puzzles

The first puzzle is Frege's puzzle of informative identity statements; how can it be that the thought *Hesperus is Hesperus* is trivial, whereas *Hesperus is Phosphorus* is informative, when *Hesperus* and *Phosphorus* have the same reference? Recanati gives the following explanation of the puzzle: It is not unusual that someone has multiple files regarding the same object. For instance, I might see someone cutting grass and fail to recognize him as Noam Chomsky. I then have one (perceptual) file referring to Chomsky and also one (conceptual) file referring to Chomsky. Likewise, for the Ancient Babylonian *Hesperus* and *Phosphorus* were distinct mental files. That is to say, all information associated with *Hesperus* is located within one file, whereas all information associated with *Phosphorus* is stored in another file. Before making the discovery, the information contained within each file is isolated from that in the other. In the case of someone thinking *Hesperus is Hesperus,* both terms are associated with the same mental file, *Hesperus,* and so the statement is trivial. Since the mental files *Hesperus* and *Phosphorus* are isolated, one is not rationally allowed to infer from this information that the same heavenly body is visible twice a day, since there is no information in the *Hesperus* file that predicates of Hesperus that it is visible any other time than during the evening. According to Recanati, what happens when one makes the discovery that Hesperus is Phosphorus is that information between the two files *Hesperus* and *Phosphorus* can flow

freely from one file to the other. This he calls *linking*: "When two files are linked, information can flow freely from one file to the other" (Recanati 2012, 43). When learning that Hesperus is Phosphorus, the Ancient Babylonians were able to use information from both files, and infer that the same heavenly body is visible twice a day.[68]

In 4.1 I argued that Sainsbury and Tye fail to explain this puzzle. The reason given was that they say that the discovery that Hesperus is Phosphorus is only a cognitive discovery. Frege rejects this view when he lays out the puzzle in his (1892). The puzzle posed by Frege is how identity statements such as *Hesperus is Phosphorus* can be an *empirical* discovery if the meaning of the concepts (terms) is exhausted by their reference. Sainsbury and Tye thus fail to explain the puzzle posed by Frege since they hold that coming to know that Hesperus is Phosphorus involves only a cognitive discovery. Recanati's solution to Frege's puzzle, on the other hand, does not face the same worries that the originalist solution did. Recanati agrees with Sainsbury and Tye that the thoughts *Hesperus is Hesperus* and *Hesperus is Phosphorus* have the same truth conditions. However, Recanati can allow for the discovery in question to involve new knowledge about the world. The mental files contain information about the object to which they stand in ER relations; the information in the mental files is information about objects in the world. On this picture, when the Ancient Babylonians learned that Hesperus is Phosphorus there was a linking between their files *Hesperus* and *Phosphorus*: After the discovery, all information regarding the reference within one of the files becomes accessible through the other. Since the information that Hesperus is visible in the evening now becomes available through the *Phosphorus*-file, and this file already contains the information that Phosphorus is visible in the morning, this explains how the Ancient Babylonians learned something new about the world; namely that the same heavenly body is

[68] On this view, linking is also what explains recognition: When I succeed to recognize the man cutting the grass as Chomsky, information flows freely from the perceptual file to my conceptual file associated with Chomsky.

visible twice a day.[69] Hence, Recanati succeeds in explaining how one can make an empirical discovery through informative identity statements. Furthermore, in 4.2 I argued that Sainsbury and Tye must stipulate that *Hesperus* and *Phosphorus* have distinct origins in order to explain the case of informative identity statements; if it had turned out that the concepts had the same origin, Sainsbury and Tye would encounter analogous problems to those posed for their solution to the puzzle of Paderewski in 4.1. Recanati, on the other hand, is in no need of any such stipulations; an appeal to the mental files being distant is sufficient for the explanation of the puzzle. Recanati takes mental files to be transparent to a thinker; one cannot be wrong about the number of vehicles involved in a thought. Hence, there is no room within his theory for the Ancient Babylonians' concepts *Hesperus* and *Phosphorus* being of the same type. It is clear then that Recanati does not face the problems posed for originalism in accounting for informative identity statements.

I now turn to the puzzle of twins; how can it be that qualitatively identical duplicates can be in qualitatively distinct mental states? I argued that Sainsbury and Tye's account does not really add anything to the classical puzzle they purport to solve. Their solution to the puzzle does nothing to accommodate the externalist intuition that even if Oscar and Twin Oscar have different mental content when thinking *water is wet,* Gell-Mann and Twin Gell-Mann have the same mental content when thinking *quarks are tiny.* I will now show that the mental file framework of Recanati can better accommodate this intuition and make a contribution to the classic debate. Unlike Sainsbury and Tye, Recanati takes meaning vehicles to be individual. That is to say, Oscar and Twin Oscar do not share haecceitistically identical mental files, since the files are located within individual minds. What is important is not that the twins

[69] It might be argued that Sainsbury and Tye could explain the Ancient Babylonians making an empirical discovery in a similar way to Recanati. Since the framework of the two theories are so similar, it might be that Recanati's explanation is also available to Sainsbury and Tye. However, Sainsbury and Tye are clear that identity statements can only involve cognitive discoveries. In fact, they must say this in order to give their specific explanation of the knowledge argument stemming from Jackson (1982). Here they defend physicalism by saying that Mary only makes a cognitive discovery when leaving her black and white room. In short: When still in the black and white room Mary knows one answer to the question as to what it is like to see red. She knows that *this* (referring to what she has come to know about the experience of seeing red through her books) is what it's like to see red. After having seen a red rose for the first time Mary comes to know a new answer to the question: she knows that *this* (referring to her current experience) is what it's like to see red. When Mary forms the thought *this is this* (where both tokens refer to what it's like to see red, and the first concept token is the concept she used before her release and the latter token is the one she acquired after her release), both concepts refer to the experience of seeing red, but since Mary introduced the indexical concepts at distinct events, they are of different types. This explains why her thought is informative. Her thought being informative, however, does not involve her making an empirical discovery: "Mary makes a discovery when she leaves the room. But if physicalism is true, her discovery is a cognitive discovery […] In this respect, it is like the discovery that Hesperus is Phosphorus or that Cicero is Tully" (Sainsbury and Tye 2012, 166). Hence, it seems that Sainsbury and Tye must say that coming to know informative identity statements only involves a cognitive discovery.

share the same meaning vehicles, but that their meaning vehicles are qualitatively identical, or – in the case of *water* and *twin water* – qualitatively distinct. Oscar and Twin Oscar's concepts *water* are qualitatively distinct, since one refers to H_2O while the other refers to XYZ. The other concepts used by the twins, however, should turn out to be qualitatively identical, since they refer to qualitatively identical objects (the only difference between Earth and Twin Earth, recall, is the chemical structure of water). Mental files, recall, are typed by their function to store information about certain objects. Consider first the case of *quarks* and *twin quarks*, discussed in 4.3: Gell-Mann and Twin Gell-Mann stand in relations to qualitatively identical objects, namely quarks, so their *quark* files function to store information about qualitatively identical objects. Since Gell-Mann and Twin Gell-Mann are qualitatively identical duplicates and their *quark* files were created in qualitatively identical contexts, the ER-relations between the files and the objects must be qualitatively identical as well. This explains why their thoughts *quarks are tiny* are qualitatively identical, since mental files are constituents of thoughts and Gell-Mann and Twin Gell-Mann have qualitatively identical files.

Now, consider the case of Oscar and Twin Oscar: The twins stand in relations to qualitatively distinct objects; H_2O and XYZ. Since their files function to store information about qualitatively distinct objects, the ER-relations are qualitatively distinct. Hence, when thinking *water is wet* Oscar and Twin Oscar activate qualitatively distinct mental files. A similar explanation may be given to Burge's thought experiment regarding actual and counterfactual Alf: In the actual case, Alf stands in an ER relation to arthritis whereas, in the counterfactual case, he stands in an ER relation to tarthritis. Since the contexts are qualitatively distinct, so are the mental files associated with each scenario. Further, since Recanati takes the reference to be determined by the ER relation, this account also explains what originalism failed to explain; how intrinsic duplicates can differ with respect to their mental content. Gell-Mann and Twin Gell-Mann stand in qualitatively identical relations to *quarks,* so the references of their *quark* files are qualitatively identical. Oscar and Twin Oscar, on the other hand, stand in qualitatively distinct ER relations to H_2O and XYZ respectively, and hence the reference of the two files are also qualitatively distinct. Hence, Recanati avoids the criticism put forth against the originalist solution to the puzzle of twins in 4.3.

The third puzzle is the puzzle of *cats* and *chats*. In solving the puzzle of Paul making a discovery when learning that *cat* and *chat* have the same reference*,* the originalists must yet

again stipulate that the concepts have distinct origins. In case of the concepts having the same origins, the originalists had to appeal to Paul being mistaken about the number of concepts deployed in thought. Recanati does not need to stipulate that the concepts have distinct origins or Paul being mistaken about the number of concepts involved. As mentioned, Recanati denies that one can be wrong about the number of files one has; he takes modes of presentation (non-semantic cognitive content) to be transparent to the thinker. This is one of Recanati's main reasons for making a distinction between mode of presentation and reference: "If modes of presentation are not transparent, there is no reason to move from pure referential talk to mode of presentation talk in the explanation of rational behaviour" (Recanati 2012, 116-117). Paul's believing that *cats* and *chats* are distinct concepts, then, shows that he has distinct mental files associated with cats. When learning that *cats* and *chats* have the same reference, there is a linking between Paul two files, allowing information to flow freely between the two files. This explains why Paul makes a discovery when learning that the concept *cat* has the same reference as the concept *chat*.

The fourth puzzle is the Puzzle of Paderewski. In 4.2 I showed that the originalist solution involved attributing an indefinite amount of contradictory *n*-order beliefs, and thus that originalists fail to explain Peter being rational. Unlike Sainsbury and Tye, Recanati takes meaning vehicles to be individual. In the case of Peter having inconsistent thoughts about Paderewski, the mental file framework can account for Peter's being rational by saying that he has distinct files that both refer to Paderewski. One of Peter's Paderewski files contains the information gathered about Paderewski at the concert, while the other contains the information gathered about Paderewski at the rally. These files are not linked, since Peter does not know that the information is about the same individual. When forming the belief *Paderewski has musical talent*, Peter uses only information from the file containing the information from the concert. When forming the belief *Paderewski does not have musical talent*, Peter only uses information from the file containing the information from the rally. This explains why Peter is rational in holding contradictory beliefs. When Peter makes the discovery *Paderewski (the musician) = Paderewski (the politician)* there is a linking between the two files which account for him learning a new fact. While originalists must say that Peter only has one *Paderewski* concept, Recanati can allow for Peter's thoughts containing distinct meaning vehicles, and thus Recanati need not appeal to any such thing as Peter having false second order beliefs in order to explain his being rational. Recanati thus avoids the problems posed for originalism in 4.2.

I now turn to the puzzle about pure demonstratives. This is the puzzle of how it is possible to fail to know whether identity statements on the form *that is that* are true, when both tokens of *that* have the same reference. In this case, Recanati can say that the individual has two indexical mental files *that*, both referring to the same object. Before actually believing that the two tokens of *that* refer to the same thing, the files are not linked, and thus information cannot flow freely between the two files. When wondering whether that is that, two distinct files are active, and this explains why learning that the thought *that is that* is true, is informative. The solution to this puzzle based on Recanati's framework bears similarity to the originalist solution to the same puzzle. This is because originalists agree with Recanati that indexical concepts are individual. Both the originalist solution to the puzzle and the solution just presented appeal to the thinker having distinct meaning vehicles when wondering if the thought *that is that* is true. I consider the originalist solution and the solution based on Recanati's framework to the puzzle of pure demonstratives to be on equal footing.

The next puzzle is the puzzle of empty thoughts. Since, according to Recanati, thought vehicles are typed by their function, and this involves acquaintance, it seems that thoughts deploying empty concepts cannot have the same status as thoughts having a reference. Recanati says that one can open a file without actually being acquainted with the relevant object, but "opening a mental file itself is *not* sufficient to entertain a singular thought (in the sense of thought-content)" (Recanati 2012, 164). Further, he says that mental file tokening "is sufficient to entertain a singular thought only in the sense of thought-*vehicle*" (Ibid., 164). That is, even though singular thought vehicles are *typed* by their function, Recanati holds that such entities can be tokened even if this function is not fulfilled. The function to store information about the referent is not a *de facto* function of the files, but rather it is a *de jure* function. This means that the mental files can exist even if they fail to fulfill their function. However, if there is no object to be acquainted with, one cannot think a thought in the ordinary sense, but instead one can entertain a thought vehicle.[70] This solution resembles the

[70] In a forthcoming article, Carsten Hansen and George Rey argue that Recanati's theory commits him to a form of *actualism*. This is the view that thought about individual objects "is thought about *actual* objects virtually all external to the cognitive system" (Hansen and Rey (forthcoming), 1). They argue that by taking singular thought to depend on acquaintance, Recanati's theory fails to give an account involving various thoughts about objects we can never be acquainted with. They suggest that the mental framework can be maintained without appealing to acquaintance: "we see Recanati's very own postulation of mental files as "senses" of tokens in contexts as an excellent suggestion for the content of empty terms – provided of course it's freed of the commitment to acquaintance with real objects!" (Hansen and Rey, 11). For present purpose I will grant Recanati's explanation of empty thoughts. It may, however, be that we would be better off if we adopt a view similar to Recanati's but that is not committed to acquaintance relations to actual objects.

one given by Sainsbury and Tye in that both theories appeal to the vehicles of content when explaining empty concepts playing an interesting role in cognition.

The final puzzle addressed by Sainsbury and Tye is the puzzle of thinking about oneself. Recanati holds that mental files possess the essential features of indexicals in that mental files are context sensitive. Regardless of context, however, one is always in an ER relation to oneself. For instance, when you leave a room you no longer think of the room as 'here', but you always think of yourself as 'I'. For Recanati, then, the *self* file is a stable file, meaning that it is not temporal in the way that most other mental files are (Recanati 2012, 68). Further, as noted, Recanati holds that (conceptual) mental files contain two sorts of information: "information gained in the special way that goes with that relation (first person information in the case of the *self* file), and information not gained in this way but *concerning the same individual as information gained in that way*" (Ibid., 65-66). For instance, knowledge of one's birthday is not information gained through the relation to oneself, but rather it is learned from the testimony of others. Since one thinks this information gained from others regards oneself, the information is stored in one's *self* file. Information gained in this way might be false; someone might have lied when informing you of your birthday. The other kind of information, the one you have in virtue of standing in a stable acquaintance relation to yourself, on the other hand, is immune to error (for instance, you cannot be wrong about whether you are in pain or not). Further, in the case of Mach thinking that the man in the mirror is a shabby pedagogue while failing to recognize that the man in the mirror is himself, the information gained when thinking about himself in third person does not go into the *self* file, since Mach does not think the information gained is about himself. When learning that the man in the mirror is himself, there is a linking between Mach's *self* file and the file containing the information about his own reflection. Recanati's solution to this puzzle is similar to the one proposed by Sainsbury and Tye in that they both claim that Ernst Mach uses distinct vehicles when thinking about himself in third- and first person perspective. I problematize neither solution in this thesis.

I have shown, very briefly, that Recanati's theory of mental files provides easy and straightforward solutions to the puzzles addressed by Sainsbury and Tye. The theory also succeeds in accounting for the problems I raised in chapter 4 for Sainsbury and Tye's originalist solutions to the puzzles. Hence, the mental file framework is preferable to Sainsbury and Tye's originalism when it comes to solving the puzzles. I will now show that

Recanati also avoids the problems raised for Sainsbury and Tye's account by my thought experiment in chapter 5.

7.3 The Thought Experiment Revisited

The problem posed for originalism in chapter 5 was that the theory allows for individuals using concept tokens of the same type to express contradictory contents. If the correct understanding of Sainsbury and Tye's account is that cognitive significance is to be explained in terms of vehicles rather than content, the theory fails to explain the cognitive significance of thoughts that are the same at the level of vehicles but contradictory at the level of semantics, and vice versa. A further consequence is that the theory fails to explain why someone is rational in accepting or rejecting deductions that are valid at the level of syntax but invalid at the level of semantics, and vice versa. This implies that the individuation of concepts and thoughts should not come apart from reference fixing if one takes cognitive processing to depend directly on the vehicles of content. As noted, it may be that even if Sainsbury and Tye explain the puzzles in terms of vehicles they are not committed to saying that *every* instance of cognitive significance is to be explained in terms of vehicles alone. By making some adjustments to my thought experiment I showed that their account fails to give a general account of rationality, even if their framework allows them to appeal directly to the content of thoughts. I will now show that Recanati does not face any such problems with the thought experiment put forth in chapter 5.

The reason why originalists must claim that $ohun_{(A)}$ and $ohun_{(B)}$ are of the same type is that the concepts have the same origin. Recanati, however, who does not take concepts to be individuated by their origin, would not have to claim that the concepts are of the same type.[71] For Indira to acquire the concept $ohun_{(A)}$ involves her standing in ER relations to blue objects. Her acquiring the concept $ohun_{(B)}$, on the other hand, involves her standing in ER relations to objects that are not blue. The information gained through these ER relations is stored in distinct mental files. In this case, $ohun_{(A)}$ and $ohun_{(B)}$ are not equivocal (as they must be on the originalist account). To Recanati, they are simply distinct concepts which names are spelled and pronounced the same way (like 'bank' in the sense of *river bank* and 'bank' in the sense

[71] Note that Recanati's theory concerns singular thoughts. Some of the thoughts I discuss here are general thoughts, however (they are not about particular objects). I propose a simple extension of Recanati's framework in order to deal with general thoughts.

of *money bank*). When thinking the thoughts *an ohun$_{(A)}$ is blue* and *an ohun$_{(B)}$ is not blue,* distinct mental files are active in Indira's mind. Hence, the two thoughts are not contradictory at the level of vehicles, since they do not deploy the same concepts. Further, in the case of Indira wanting a blue bike and forming the thought *that bike is an ohun$_{(A)}$* she would be rational in buying the bike, whereas had she instead formed the thought *that bike is an ohun$_{(B)}$* she would not be rational in buying the bike. Hence, if Indira is rational, the two thoughts would have different causal powers. Sainsbury and Tye, recall, could not explain the difference in causal powers of the two thoughts, since they must say that the thoughts are of the same type. Recanati, on the other hand, has no problem explaining the cognitive difference of the two thoughts. Since, on this view *ohun$_{(A)}$* and *ohun$_{(B)}$* are distinct concepts, the thoughts are of different types. The thoughts' being of different types explains why they play different roles in cognition. The thoughts playing different roles in cognition explains why the two thoughts have different causal powers. Hence, Recanati's theory provides a straightforward explanation of the difference in the two thoughts in terms of vehicles of content.

Further, in the case of Indira performing deductions in her head, the difference in reference is reflected in a difference in vehicles, on the mental file framework. Recanati says that "unicity of reference is a built-in presupposition of the file" (Recanati 2012, 132). That is to say, someone cannot knowingly use one file to refer to distinct objects.[72] Consider for instance the case of Indira performing the following deduction in thought, addressed in chapter 5:

(2) P1: If something is an ohun$_{(A)}$ then it is blue

P2: The book is an ohun$_{(B)}$

C: The book is blue

To Recanati the deduction would look the same regardless of whether it is formulated with respect to vehicles or reference. If considering only the vehicles, the deduction is invalid - so Indira is rational in rejecting the deduction. This is the desired result. The problem with the deduction only rises if one takes *ohun$_{(A)}$* and *ohun$_{(B)}$* to be of the same type. Sainsbury and

[72] In the case of someone unknowingly storing information about distinct objects (as in inverse Paderewski cases) Recanati says that the file fails to refer: "If there is no object, or more than one, the file does not refer" (Recanati 2012, 132). Hence, one and the same mental file can never have two distinct references, on Recanati's view.

Tye hold that "both being contradictory and being valid essentially depend on how many concepts are involved" (Sainsbury and Tye CC, 3-4). In chapter 5 I showed that if allowing for concepts of the same type to express distinct content, this statement fails. However, if one holds that a difference in reference entails a difference in concepts, this statement holds.

Further, Recanati holds that concepts are individual rather than public. If confronted with the thought experiment as formulated in 5.3, he would encounter no problems in accounting for Indira being rational. On Recanati's framework, the reference of the vehicles is whatever they stand in ER relations to. Even though the other people in both group A and group B have started to use the concept *ohun* to pick out anything that is an object, the reference of Indira's files $ohun_{(A)}$ and $ohun_{(B)}$ still have the same reference as before she left, since the files stand in ER relations to objects that are blue and objects that are not blue, respectively. When Indira forms the thought *an* $ohun_{(A)}$ *is blue* this thought is true since the predicate 'is blue' is stored within a file referring to blue objects. Likewise, her thought *an* $ohun_{(B)}$ *is not blue* is also true, since the predicate 'is not blue' is stored within a file referring to objects that are not blue. Since Indira's concepts $ohun_{(A)}$ and $ohun_{(B)}$ differ both at the level of vehicles and reference, the two thoughts are not contradictory. Since the two thoughts are not contradictory there is nothing threatening Indira being rational on this view. Hence, Recanati's theory is apt to explain both versions of the thought experiment put forth in chapter 5. On the basis of the cases discussed in this thesis, then, we have reasons to prefer Recanati's theory of mental files to Sainsbury and Tye's originalist account.

7.4 Chapter Summary

In this chapter I've presented an alternative to originalism: Recanati's theory of mental files. The purpose of this chapter has been to show that one can agree with originalism that the explanation of cognitive significance does not depend on a semantic notion of mode of presentation but rather on the vehicles of content, and at the same time avoid the problems posed for originalism in this thesis. In particular, I showed that Recanati's framework provides solutions to the seven puzzles of thought – also those that I argued originalism fails to solve. In particular, Recanati can account for informative identity statements in terms of vehicles without making any stipulations about the origins of such vehicles. When the Ancient Babylonian discovered that *Hesperus is Phosphorus* the two files were linked and

information could then flow freely between the two files. This explains why the Ancient Babylonians made an important empirical discovery when learning that Hesperus is Phosphorus; after the discovery, all information regarding the referent of one file could be accessed through the other. Recanati's framework also provides an unproblematic explanation of the puzzle of Paderewski; Peter has two distant files, that unbeknownst to him refer to the same individual. Since he uses one file when thinking that Paderewski has musical talent and another when thinking that Paderewski lacks musical talent, this explains him being rational. Thus, Recanati does not need to appeal to Peter having false higher order beliefs, as did Sainsbury and Tye. Finally, Recanati's framework also explains the difference in thought of Oscar and Twin Oscar; since Oscar stands in an ER relation to H_2O and Twin Oscar stands in a relation to XYZ, their mental files are qualitatively distinct. Gell-Mann and Twin Gell-Mann, on the other hand, stand in ER relations to qualitatively identical objects, so their mental files are qualitatively identical. Since the ER relation also determines the reference of the files, this explanation also contributes to the classic debate about mental content. Hence, Recanati's framework succeeds in explaining the puzzles originalism fails to solve.

Furthermore, I argued that Recanati avoids the problems posed for originalism by the thought experiment presented in chapter 5. This is because, unlike originalism, Recanati's framework does not allow for one vehicle to have more than one reference; a difference in reference indicates a difference in files. As I have stressed, this is not intended as a general argument in favour of Recanati's theory – there might be other problems with his theory. Rather, what this shows is that, on the basis of the cases discussed in this thesis, we have reasons to prefer Recanati's theory of mental files to originalism as stated in *Seven Puzzles of Thought*.

Chapter 8

Conclusion

I have argued that originalism, as stated in *Seven Puzzles of Thought*, fails as a general theory of concepts and thoughts. In chapter 4 I showed that originalism fails to solve certain of the puzzles the theory is advanced to solve. In particular, I argued that their solution to the puzzle of Paderewski fails, since the appeal to Peter having false second order beliefs leads to an infinite regress of more sets of contradictory beliefs at increasing metalevels. If anything, then, Peter's being rational is rendered even harder to explain after making the appeal to false higher order beliefs, since we end up stipulating him having more sets of contradictory beliefs than he had in Kripke's original puzzle. I then argued that the originalist solution to the puzzle of *Hesperus* and *Phosphorus* fails to explain how identity statements can potentially provide new knowledge about the world. The reason for this is that, according to originalism, when learning that Hesperus is Phosphorus, the Ancient Babylonians made only a cognitive discovery and not an empirical discovery. Originalism thus fails to explain what was at the centre of Frege's original puzzle. My final criticism of the originalist solution to the puzzles concerned the solution to the puzzle of twins. I argued that Sainsbury and Tye do not correctly represent the puzzle rising from the classical debate about mental content stemming from Putnam and Burge. Their solution to the original puzzle is irrelevant, and their solution to the puzzle they draw from the original puzzle (how intrinsic duplicates may differ with respect to their thoughts (vehicles of content)) depends on serious stipulations and fails to reflect the difference between intrinsic duplicates having the same mental states on the one hand and their having distinct mental states on the other. The fact that Sainsbury and Tye's account offers no new contribution to the questions that puzzled Putnam and Burge, is reflected in their claim that the core assumptions of originalism are consistent with both internalism and externalism about mental content. Oscar and Twin Oscar's concepts *water*

would be distinct, and thus play different roles in cognition, even if the watery stuff on Twin Earth had the same chemical structure as the watery stuff on Earth. In *Seven Puzzles of Thought* the reason given for preferring originalism to competing theories is that the theory solves the classical puzzles. Since originalism fails to solve three of the classical puzzles it sets out to solve we are given little reason to prefer originalism to competitive theories.

Since Sainsbury and Tye's main focus is on the seven puzzles of thought, they give no explicit statement of what a general theory of cognitive significance would be on their account. What is certain is that their only explanatory tools are vehicles of content and reference; in particular, their explanatory framework does not appeal to any notion of cognitive content. What is unclear is whether Sainsbury and Tye can appeal directly to the reference in cases where the cognitive significance cannot be explained in terms of vehicles. In order to show that their theory fails as a general theory regardless of their take on the explanatory role of reference, I have proposed two versions of a thought experiment. In the first version of the thought experiment, I argued that if Sainsbury and Tye take cognitive significance to be explained purely in terms of vehicle of content rather than the content expressed by such entities, they fail to explain how someone can be rational in holding beliefs that are contradictory at the level of syntax, but not at the level of content. Also, since they allow for concepts to be equivocal, they cannot explain how someone can be rational in accepting deductions that are invalid at the level of syntax but valid at the level of semantics, if taking cognitive significance to depend solely on the vehicles of content. In the second version of the thought experiment, I argued that even if Sainsbury and Tye are not committed to saying that every instance of cognitive significance is to be explained by appeal to vehicles alone (i.e. if they can allow the explanation to depend directly on the semantic features of thoughts), they still fail to explain certain cases of someone being rational in holding seemingly contradictory beliefs. In both formulations of the thought experiment, the problems posed for Sainsbury and Tye are due to their commitment that a change in reference does not necessitate a change in vehicles. However, since one of the key claims of originalism is that concepts are to be individuated by their origins and not by their semantic properties, Sainsbury and Tye must allow for concepts expressing contradictory content on their current framework. Since their theory fails to explain certain cases of rationality, then, Sainsbury and Tye's originalist account fails as a general theory of concepts and thoughts.

I have argued that there are other theories, such as Recanati's theory of mental files, that are better suited to solve the puzzles than originalism, and also that such theories also avoid the problems posed for originalism by my thought experiment. Recanati's theory takes concepts to be individuated in terms of their function, and their function is to store information about the objects to which they stand in epistemically rewarding relations. Hence, the individuation of concepts depends on their relation to their specific referents. Therefore, this theory cannot allow for a change in reference without a change in vehicles. I have indicated that this last commitment is one that we should endorse in giving a general theory of concepts and thoughts, if we are to explain certain cases of rationality. There might be other problems the originalist account may avoid which confront theories that have this commitment, but on the basis of the cases discussed in this thesis, we have reasons to reject Sainsbury and Tye's originalism in favour of competing theories of concepts and thoughts.

Literature

Austin, David. 1990. *What is the Meaning of "This"? A Puzzle about Demonstrative Belief*. New York: Cornell University Press.

Burge, Tyler. 1979a. "Sinning Against Frege". Reprinted in his 2005.

_____.1979b. "Individualism and the Mental," in French, Uehling, and Wettstein (eds.), *Midwest Studies in Philosophy*, IV, Minneapolis: University of Minnesota Press, pp. 73–121.

_____. 1988. "Individualism and Self-Knowledge," *Journal of Philosophy*, 85(1): 649–663

_____. 2005. *Truth, Thought, Reason: Essays on Frege*, Oxford: Oxford University Press

Campbell, John. 2002. *Reference and Consciousness,* Oxford: Oxford University Press.

Chalmers, David. 2004. "Epistemic Two-Dimensional Semantics", *Philosophical Studies*, 118: 153–226.

_____. 2006. "Two-Dimensional Semantics", in E. Lepore and B. Smith (eds.), *Oxford Handbook of Philosophy of Language*, Oxford: Oxford University Press, pp. 575–606.

Devitt, Michael. 1981. *Designation*, New York: Columbia University Press.

Evans, Gareth. 1973. "The Causal Theory of Names," *Proceedings of the Aristotelian Society*, Supplementary Volume 47: 187–208.

Fodor, Jerry. 1975. *The Language of Thought*, Cambridge, Massachusetts: Harvard University Press.

_____. 1980. “Methodological solipsism considered as a research strategy in cognitive psychology”, *Behavioral and Brain Sciences* (3), No.1: 63- 73

_____. 1994. *The Elm and the Expert*, Cambridge, Massachusetts: The MIT Press.

_____. 2008. *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press

Frege, Gottlob. 1879. *Begriffsschrift,* chapter 1 in P. Geach & M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege* (1960), Oxford: Blackwell.

_____. 1892. “On Sense and Reference”*,* in P. Geach & M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege* (1960), Oxford: Blackwell.

Gerken, Mikkel. 2013. *Epistemic Reasoning and the Mental*. Palgrave Macmillan

Hedger, Joseph. Forthcoming. “Sainsbury and Tye Fail to Solve Frege’s Puzzle”, in *Linguistic and Philosophical Investigations.*

Horwich, Paul. 2014. “Critical Notice of Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts, by R. M. Sainsbury and Michael Tye”, *Mind* 123 (492): 1123—39.

Jackson, Frank. 1982. “Epiphenomenal Qualia”, *Philosophical Quarterly* 32: 127–136.

Kallestrup, Jesper. 2012. *Semantic Externalism.* London: Routledge.

Kripke, Saul. 1980. *Naming and Necessity,* Cambridge, MA: Harvard University Press.

_____. 1979. “A Puzzle about Belief” in A. Margalit (ed.), *Meaning and Use*. Reidel. (239—83).

Lewis, David. 1986. *On the Plurality of Worlds*, Oxford: Blackwell Publishers.

Loar, Brian. 1987. “Social Content and Psychological Content”, in R. H. Grimm & D. D. Merrill (eds.), *Content of Thought.* Tucson: University of Arizona Press: 99-110.

Mach, Ernst. 1914. *The Analysis of Sensations*. London: Open Court Publishing Company.

McKay, Thomas. 1984. “Critical Review of Michael Devitt's *Designation*”, Noûs 18, 357-367.

Mill, John Stuart. 1843. *A System of Logic,* London: Longmans.

Millikan, Ruth. 2011. “Losing the Word-Concept Tie.” *Proceedings of the Aristotelian Society,* Supp. 111: 125—43.

Mole, Christopher. 2013. "Attention", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2013/entries/attention/ [accessed 21.04.2015]

Putnam, Hilary. 1975. “The Meaning of ‘Meaning’” in A. Pessin & S. Goldberg (eds.), *The Twin Earth Chronicles* (1996), Armonk, NY: M.E. Sharpe.

Recanati, François. 2012. *Mental Files.* Oxford: Oxford University Press.

Rey, Georges. 1997. *Contemporary Philosophy of Mind: a Contentiously Classical Approach*, Oxford: Blackwell

Rey, Georges and Carsten Hansen. Forthcoming. "Files and Singular Thoughts Without Objects or Acquaintance: The Prospects of Recanati's (and Others') "'Actualism"' in *Review of Philosophy and Psychology* (expected Fall 2015).

Russell, Bertrand. 1911. “Knowledge by Acquaintance and Knowledge by Description,” *Proceedings of the Aristotelian Society*, 11: 108-128.

Sainsbury, Mark and Michael Tye. 2011. “An Originalist Theory of Concepts”. *Proceedings of the Aristotelian Society,* Supp. Vol. 85: 101—24.

_____. 2012. *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts.* Oxford: Oxford University Press.

_____. Forthcoming. “Counting Concepts: Response to Boghossian,” in S. Goldberg (ed.) *Externalism and Skepticism*, Cambridge University Press.

Schiffer, Stephen. 1992. “Belief Ascription,” *Journal of Philosophy* 89: 499-521.

Searle, John. 1958. “Proper Names”, *Mind*, 67(266): 166–73.

_____. 1983. *Intentionality*, Cambridge: Cambridge University Press.

Strawson, Peter. 1959. *Individuals: an Essay on Descriptive Metaphysics*, London: Methuen.

Wittgenstein, Ludwig. 1958. *The Blue and Brown Books*, Oxford: Blackwell.